PCT/GB 00/ 03156

10/069646

INVESTOR IN PEOPLE

CERTIFIED COPY OF PRIORITY DOCUMENT

PRIORITY DOCUMENT

SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

The Patent Office
Concept House
Cardiff Road
Newport
South Wales
NP10 8QQ

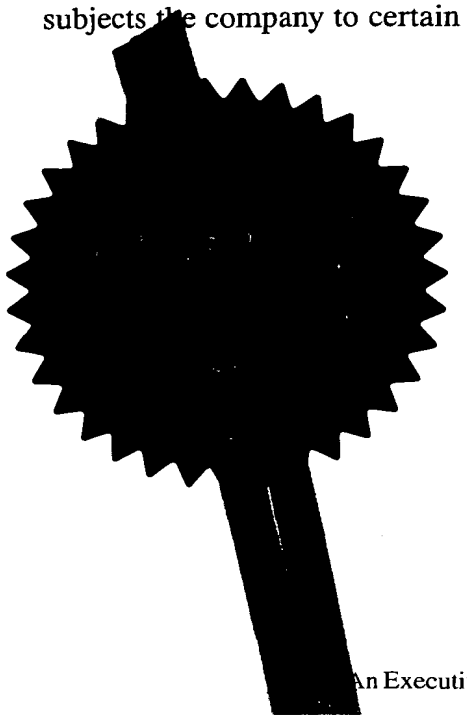
GB 00/03156

I, the undersigned, being an officer duly authorised in accordance with Section 74(1) and (4) of the Deregulation & Contracting Out Act 1994, to sign and issue certificates on behalf of the Comptroller-General, hereby certify that annexed hereto is a true copy of the documents as originally filed in connection with the patent application identified therein.

In accordance with the Patents (Companies Re-registration) Rules 1982, if a company named in this certificate and any accompanying documents has re-registered under the Companies Act 1980 with the same name as that with which it was registered immediately before re-registration save for the substitution as, or inclusion as, the last part of the name of the words "public limited company" or their equivalents in Welsh, references to the name of the company in this certificate and any accompanying documents shall be treated as references to the name with which it is so re-registered.

In accordance with the rules, the words "public limited company" may be replaced by p.l.c., plc, P.L.C. or PLC.

Re-registration under the Companies Act does not constitute a new legal entity but merely subjects the company to certain additional company law rules.



Signed J Evens.

Dated 07 September 2000

Patents Form 1/77

(Rule 16)

THE PATENT OFFICE
G
21 AUG 1999
RECEIVED BY POST

The Patent Office

21AUG99 E471328-1 000554
P01/7700 0.00 - 9919805.3

Request for grant of a patent

(see the notes on the back of this form. You can also get an explanatory leaflet from the Patent Office to help you fill in this form)

21 AUG 1999 The Patent Office

Cardiff Road
Newport
Gwent NP9 1RH

1.	Your reference	MH/M088469PGB
2.	Patent application number *(The Patent Office will fill in this part)*	**9919805.3**
3.	Full name, address and postcode of the or of each applicant *(underline all surnames)*	THE UNIVERSITY OF MANCHESTER INSTITUTE OF SCIENCE & TECHNOLOGY PO BOX 88 MANCHESTER M60 1QD
	Patents ADP number *(if you know it)*	
	If the applicant is a corporate body, give the country/state of its incorporation	UNITED KINGDOM 7737620001
4.	Title of the invention	VIDEO CODING
5.	Name of your agent *(if you have one)*	Marks & Clerk
	"Address for service" in the United Kingdom to which all correspondence should be sent *(including the postcode)*	Sussex House 83-85 Mosley Street Manchester M2 3LG
	Patents ADP number *(if you know it)*	18004

6.	If you are declaring priority from one or more earlier patent applications, give the country and the date of filing of the or of each of these earlier applications and *(if you know it)* the or each application number	Country	Priority application number *(if you know it)*	Date of filing *(day/month/year)*
7.	If this application is divided or otherwise derived from an earlier UK application, give the number and the filing date of the earlier application	Number of earlier application		Date of filing *(day/month/year)*
8.	Is a statement of Inventorship and of right to grant of a patent required in support of this request? *(Answer 'Yes' if:* *a) any applicant named in part 3 is not an inventor, or* *b) there is an inventor who is not named as an applicant, or* *c) any named applicant is a corporate body.* *See note (d))*	YES		

Patents Form 1/77

Enter the number of sheets for any of the following items you are filing with this form. Do not count copies of the same document

Continuation sheets of this form	-
Description	44
Claim(s)	-
Abstract	-
Drawing(s)	-

10. If you are also filing any of the following, state how many against each item.

Priority documents	
Translations of priority documents	
Statement of Inventorship and right to grant of a patent *(Patents Form 7/77)*	
Request for preliminary examination and search *(Patents Form 9/77)*	
Request for substantive examination *(Patents Form 10/77)*	
Any other documents *(Please specify)*	

11. I/We request the grant of a patent on the basis of this application.

Signature..... Date 20/08/99
MARKS & CLERK

12. Name and daytime telephone number of person to contact in the United Kingdom — MR M.P. HOLMES – 0161 236 2275

Warning

After an application for a patent has been filed, the Comptroller of the Patent Office will consider whether publication or communication of the invention should be prohibited or restricted under Section 22 of the Patents Act 1977. You will be informed if it is necessary to prohibit or restrict your invention in this way. Furthermore, if you live in the United Kingdom, Section 23 of the Patents Act 1977 stops you from applying for a patent abroad without first getting written permission from the Patent Office unless an application has been filed at least 6 weeks beforehand in the United Kingdom for a patent for the same invention and either no direction prohibiting publication or communication has been given, or any such direction has been revoked.

Notes

a) If you need help to fill in this form or you have any questions, please contact the Patent Office on 0645 500505.
b) Write your answers in capital letters using black ink or you may type them.
c) If there is not enough space for all the relevant details on any part of this form, please continue on a separate sheet of paper and write "see continuation sheet" in the relevant part(s). Any continuation sheet should be attached to this form.
d) If you have answered 'Yes' Patents Form 7/77 will need to be filed.
e) Once you have filled in the form you must remember to sign and date it.
f) For details of the fee and ways to pay please contact the Patent Office.

VIDEO CODING

1 Invention Claims

1. Incorporation of a psychovisual model into video coding algorithms for low bit rate applications with fixed bits per frame.
2. Introducing a coding order that reflects an approximation to the colour, multiresolution and foveated sensitivities of the human visual system (HVS).
3. Centrally weighted coding order for difference coefficient vectors and motion compensated vectors.
4. Zero vector run length encoding sequenced by the coding order of both difference coefficient vectors and motion compensated vectors.
5. An optimum decision algorithm for maximising bit usage of motion compensated vectors.

2 Background

The primary necessity for video coding algorithms is restricted bandwidth in communications applications and limited storage capacity in database applications. A standard QCIF colour video stream at 25 frames/s (fps) with 8 bits/pixel (bpp) requires a communications bandwidth of ≈ 14.5M bits/s and a one hour recording would require ≈ 6.4G Bytes of storage. Low bandwidth digital communication channels may be defined as those that operate below 64k bits/s (i.e., one ISDN basic rate channel). Therefore other typical bandwidths to consider are 28.8k bits/s for modems and ≤ 14.4k bits/s for mobile telephony applications. The problem is bound by the average compression ratio of the source video data required for the given channel capacity. For a 64k bits/s channel the stated example requires an average compression ratio of 232:1, a 28.8k bits/s channel requires 528:1 and a 10k bits/s channel, 1485:1.

Consider the video telephony application at very low bit rates (≤ 10k bits/s). The audio bit stream has the highest priority and therefore the video must be able to remain in synchrony, at least perceptually. For ease of 'lip synchronisation' and live operation there must be a minimum coding delay (≤ 100 ms) which excludes the use of source coding channel buffers and INTRA start frames. To provide reasonably smooth 'lip movement' with the audio, the frame rate should be at least 10 frames/s and should remain constant. These constraints impose a constant number of bits per frame that must be able to be changed at one frame's notice to provide for channel bit-rate fall-back capabilities. Note that the ITU-T Recommendation H.263 (H263 algorithm) designed for video telephony applications, is not well suited to these specifications.

2.1 Basic Algorithm

A block diagram of the basic algorithm without motion compensation is shown in Figure 1.

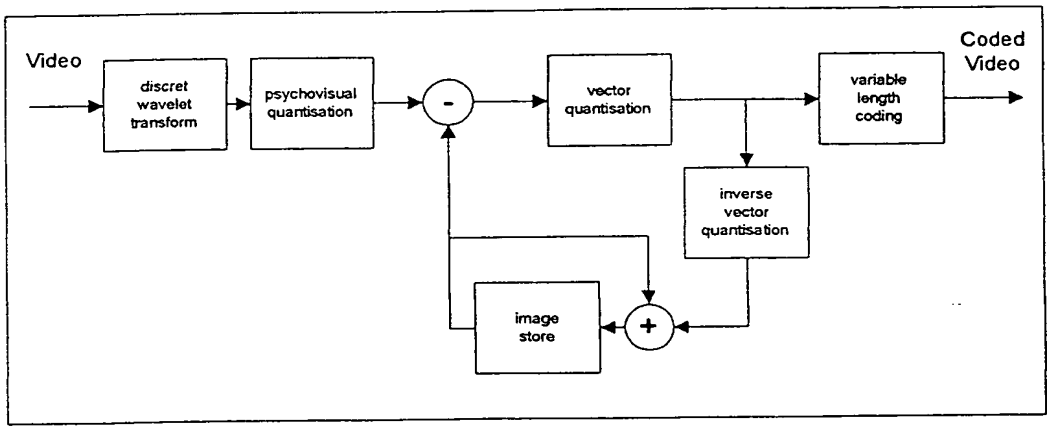


Figure 1: Block Diagram of Basic Video Coding Algorithm

For any scheme of this nature, maximum information exploitation where the output is entropy coded is achieved when the result of the information difference calculation produces small similar valued parameters. From a practical perspective this means mostly zeros and, if not zero, as close to zero as possible. Therefore the coding efficiency is focussed on the input to the vector quantiser. The difference calculation is 'encouraged' to produce small values by firstly, stripping as much spatially redundant information from the input image as possible and then secondly, to subtract from it a best possible prediction of that input.

The fundamental underlying structure of this algorithm attempts to process the video data in a manner similar to that of the human visual system. In this way the algorithm is conducive to maximising the exploitation of the subjective information redundancy. The eye is equiped to

be more sensitive to luminance information than chrominance. Therefore the video input data is represented by separate luminance (Y) and chrominance (U and V) components where the chrominance spatial resolution is reduced by two. The nomenclature for this colour space will be represented by the component's ratio as YUV 4:1:1. Colour separation and resolution reduction is a widely accepted front-end image compression method.

The early visual cortex is responsible for the non-attentive processing of the visual information presented to it by the eye's sensory mechanism. The visual information is separated into textures and edges of various orientations and processed at differing resolutions. Therefore the multiresolution wavelet transform is chosen as an appropriate domain to operate within to simulate the way in which this non-attentive processing is performed. The sensitivity to the separated sub-bands is not constant and a model of this function is used to scalar quantise the wavelet coefficients.

The wavelet transform is implemented as a discrete wavelet transform (DWT) sub-dividing the image into maximally sampled octave sub-bands. From an objective information viewpoint the choice of the particular biorthogonal filter coefficients is crucial to producing image decompositions with the minimum number of significant valued coefficients. Naturally occurring continuous-tone images are mostly composed of large smooth areas with sharp edged boundaries. The visual data, particularly at edges, is not well suited to fixed resolution harmonic function decompositions in that they produce many significant valued coefficients. Therefore an appropriately chosen DWT will, in general, perform better than harmonic transforms, such as the discrete cosine transform (DCT), in a video coding environment.

The input image with the modelled subjective information removed has the stored predicted reference DWT image subtracted from it on a coefficient by coefficient basis. A set of sub-bands and vector quantisers, one for each level, orientation and colour component, is used to quantise the difference coefficient image. The bit rate is partially controlled by an algorithm parameter that thresholds the vector coefficients before quantisation. The codebook indices are entropy coded and therefore the output bit rate is dependent on the operational parameters of the vector quantiser. The operationally optimal vector quantiser for each sub-band is found by selecting the most suitable operating point on the distortion-rate function for that sub-band. Suitability is determined from both sub-space coverage and a practical entropy code perspective.

Entropy coded codebook indices generate variable length output bit streams that are dependent on the amount of change in information between images in the video sequence.

One of the desired properties of an algorithm for video telephony applications is a constant bit rate per frame regardless of the information content. Therefore the coded bits must be allocated from the most important to the least important coefficients from a subjective viewpoint. The proposed algorithm incorporates two cognitive factors into the coding process. Firstly, objects are recognised as a whole before filling in the finer detail. In video conferencing, a human head and shoulders are noted before attempting to recognise the individual. The low resolution comes before the high and therefore is considered as more important for the algorithm. Secondly, human vision tends to be foveated especially while tracking a moving object. For video telephony, it is most likely that the face will be focussed on and will usually be in the centre of the image. Therefore it is most profitable for the coding to be centre biased. The algorithm combines these factors by coding each difference frame vector from the lowest resolution DWT sub-band to the highest and spirally from the centre outwards within the sub-band. The luminance colour components are favoured over the chrominance. The process is diagrammatically illustrated in Figure 2.

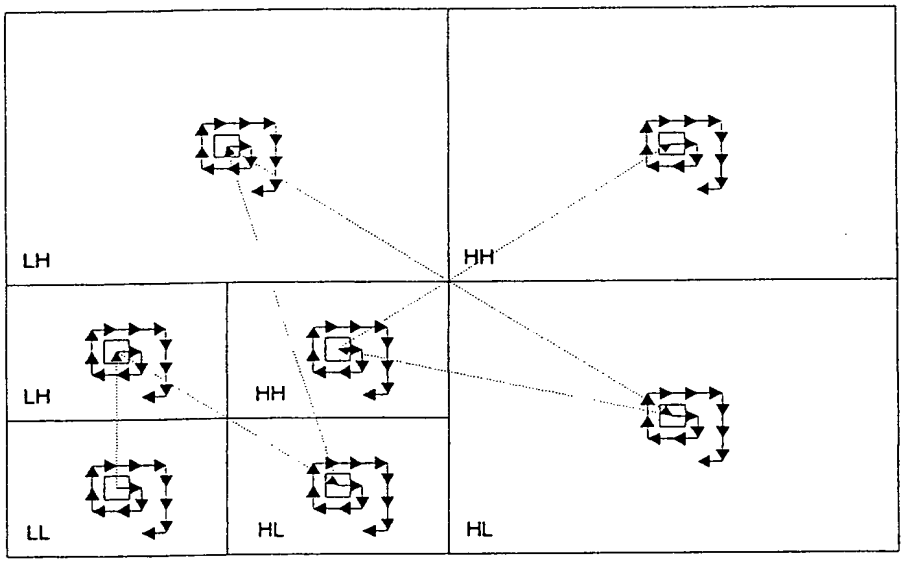


Figure 2: Foveated Coding Order for a Two Level DWT Decomposition

The majority of vectors contain all zero values therefore the coding proceeds in the ordering given by counting the zero vectors until a non-zero vector is encountered. The run of zero vectors and the codebook index of the non-zero vector are entropy coded with their respective codes. If the number of bits generated from this pair added to the current running total of bits is less than the frame bit limit then, the two entropy codes are added to the bit stream. This process continues until the frame bit limit is reached and no further bits are transmitted to the receiver. All remaining vectors are assumed to be zero by the decoder. 'End-of-sub-band' and

'end-of-image' markers are included for bit stream synchronisation. Note that the 'end-of-sub-band' markers permit the receiver to partially decode the bit stream such that a 'thumbnail' representation (i.e., lower resolution) may be viewed during reception. This is particularly useful for monitoring communication sessions or scanning stored video databases.

The artefacts generated from this coding method are a result of producing a variable length of coded vectors per frame in the ordered sequence. During periods of low temporal activity the ordered vector sequence will reach the high resolution sub-bands for the given frame rate. If this is followed by a burst of temporal activity then, the ordered sequence will only reach lower resolution sub-bands for the same frame bits and implicitly code zero change in the previously reached higher sub-bands. The visual effect is to leave motionless high frequency edge components superimposed on the moving image. This artefact is referred to as the 'shower door' effect. To alleviate the effect, the remaining higher resolution sub-band vectors in the decoded image are set to zero from the centre out until the same foveated point reached by the last coded sub-band. To allow for differing vector dimensions in each sub-band the foveated point is calculated as a fraction of the sub-band ordered sequence.

2.2 Motion Compensation

To further reduce the energy of the information presented to the vector quantiser, it is possible to track spatial translational motion between images in the video sequence. Current international standards, for example the H263 algorithm, achieve most of their compression gain from the motion compensation process in the spatial domain. It is possible to apply similar block motion estimation techniques in the DWT domain but the performance of the approximation is limited. DWT domain motion compensation has some advantages over the spatial domain counterpart in terms of the visual artefacts generated in those cases where the approximation error is large. Spatial domain compensation produces annoying blocking effects, whereas the speckled noise produced from the DWT domain compensation is less objectionable.

The basic algorithm is extended to include DWT domain motion estimation and compensation. The complete block diagram of the algorithm is shown in Figure 3.

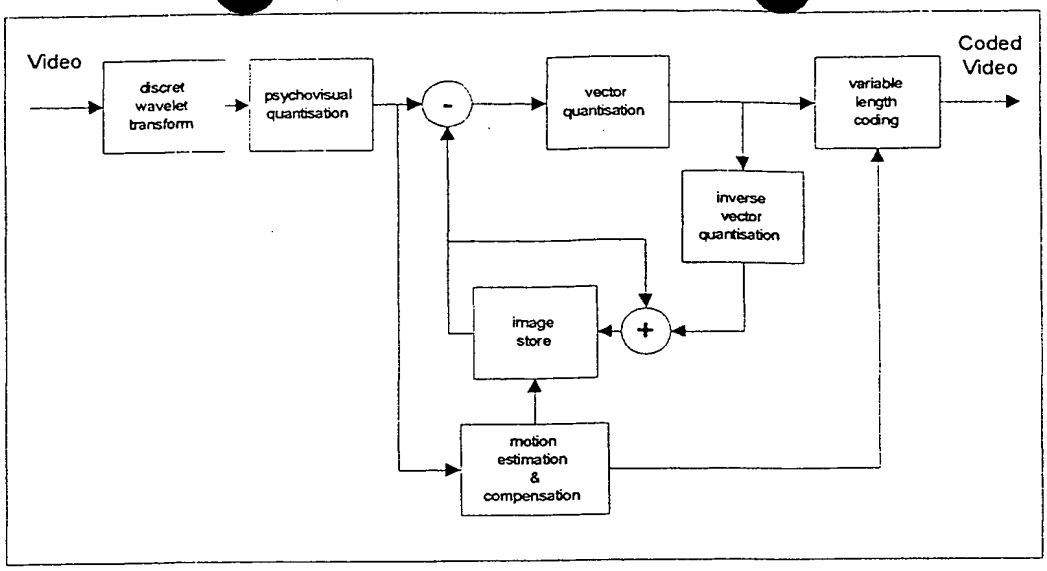


Figure 3: Block Diagram of DWT Domain Video Coding Algorithm

The basic principle of the DWT domain motion estimation and compensation proceeds in the following manner. After the subjective information has been removed from the current image, motion is estimated with each two-dimensional block of spatially corresponding coefficients within each sub-band and colour component of the stored reference image. Within each sub-band and the block dimension is chosen to be the same as that of the corresponding vector quantiser, to allow the foveated sequence to proceed on a block-by-block basis without overlap. Extending the sub-band boundaries with zero valued coefficients permits motion vectors from outside the sub-band. This increases the estimation accuracy for the boundary blocks. Half-pixel estimation is performed and the best matching block (typically in the mean –square-error sense) is written into the reference image. The compensated reference image is then subtracted from the input image and vector quantised.

There is generally little motion between consecutive frames in a video sequence particularly in the background. Therefore zero vectors will be statistically most likely. The motion vector foveated sequence is coded in a similar manner to the run length encoding of zero vectors applied to the indices of the vector quantisers. For each non-zero motion vector in the sequence, a zero run length and the two motion vector components are entropy coded and added to the output bit stream provided the frame bit limit is not exceeded.

The zero run length encoding of the motion vectors is particularly important to very low bit rate video coding algorithms where the added bit overhead may offset the quality gained by the process. For small movement within video scenes, as possible in video telephony, the

many zero and small valued motion vectors will consume scarce bit resources. An added problem with block motion estimation is that it is possible for the compensated block to produce greater difference image energy than without compensation. An uncompensated reference image block is equivalent to a compensated block with a zero motion vector. Therefore to improve the predicted reference image and to 'encourage' the zero motion vector for coding efficiency, a 'best choice' algorithm is used.

The 'best choice' algorithm is in a block mean-square-error sense. The basis of the choice is determined by vector quantising the difference image blocks from both the compensated and uncompensated images and choosing that with the lowest quantisation error. If the uncompensated image block is chosen then, the zero vector is associated with it. The process is diagrammatically illustrated in Figure 4.

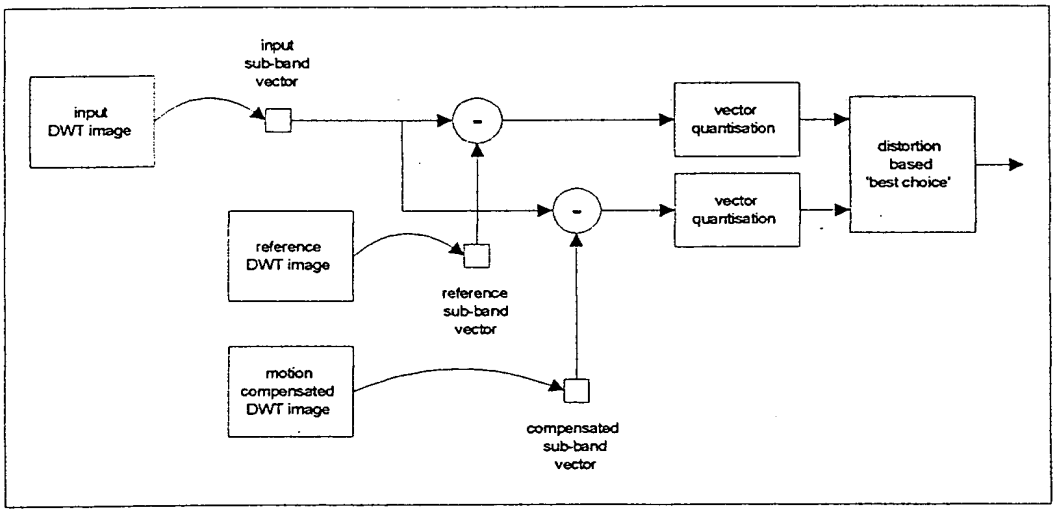


Figure 4: Block Diagram of the 'Best Choice' Algorithm

The decoder does not require any knowledge of the choice since the zero vector implicitly refers to an uncompensated block.

The extended form of the algorithm that includes motion compensation provides large performance gains during periods of large temporal movement; but even in very low bit rate environments with low temporal activity, the 'best choice' algorithm with run length coding of the zero motion vectors ensures a minimal bit cost.

3 Remarks

Although the spatial frequency and foveated psychovisual model are incorporated into a DWT domain algorithm, the technique is not bound to this type of compression. The same approach may be applied to any video coding algorithm that makes use of spatio-frequency transforms and requires a fixed number of frame bits.

In the 8x8 DCT based algorithms, the multiresolution is represented by grouping adjacent frequency coefficients within the 8x8 block into vectors. The centre 8x8 block of the difference image is then the first block in the centrally weighted coding order. DWT domain algorithms are merely more conducive to the incorporation of the proposed scheme.

oOo

Further details of the invention are given in the following appendix.

APPENDIX

1 Introduction

The primary necessity for video coding algorithms is restricted bandwidth in communications applications and limited storage capacity in database applications. A QCIF colour video stream at 25 frames/s (fps) with 8 bits/pixel (bpp) requires a communications bandwidth of ≈14.5M bits/s and a 1 hour recording would require ≈6.4G Bytes of storage. Low bandwidth digital communication channels may be defined as those that operate below 64k bits/s (one ISDN basic rate channel). Therefore other typical bandwidths to consider are 28.8k bits/s for modems and < 14.4k bits/s for mobile telephony applications. The problem is bound by the average compression ratio of the source video data required for the given channel capacity. For a 64k bits/s channel the stated example requires an average compression ratio of 232:1, a 28.8k bits/s channel requires 528:1 and a 10k bits/s channel, 1485:1.

The video coding problem may be defined as *the optimal exploitation of the subjective and objective information redundancy of a given video data representation to achieve a target short-term average channel bit rate such that the decoded video is perceptually acceptable.*

The difficulty with this problem statement is that it consists of both parts that may be mathematically defined or at least modelled with a good approximation, and parts that have no strict definition in that they are dependent on the contextual perception of the scene. An example of contextual perception (selective impairment) is the visual tolerance of a moving train with severe distortion but the same distortion of a human face is unacceptable. However, to a 'train spotter' the distorted train scene may be equally unacceptable. It may be difficult to rigorously define the *optimal exploitation*, but the engineering problem of achieving a result for given bit rates remains, albeit ill posed.

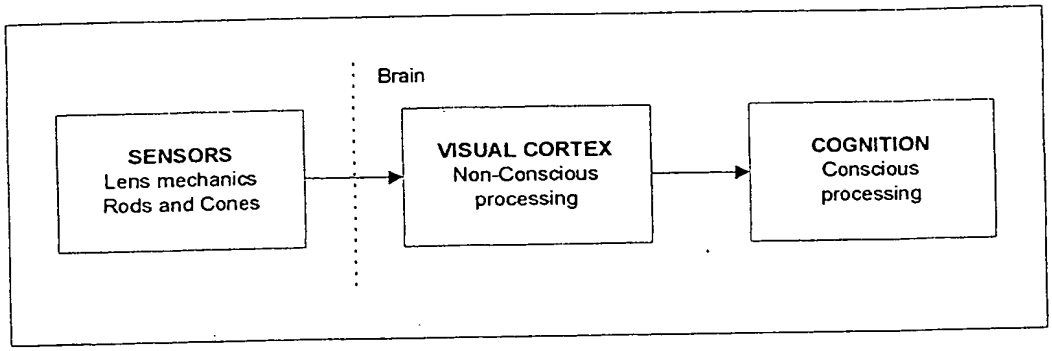


Figure 1: Model of the Human Visual System

Consider the simple model of the human visual system in Figure 1. The *subjective information redundancy* refers to that information which the visual cortex does not respond to. These are defined by the non-conscious processing of the brain such as, edge and texture separation, the frequency sensitivity and masking effects. The term non-conscious is used here to differentiate it from the Freudian unconscious terminology. Approximate mathematical models for simulating this processing have been derived empirically and may be incorporated into coding algorithms (van den Branden Lambrecht and Verscheure 1996). On the other hand, the *perceptual acceptability* of the decoded video is dependent on the cognitive functionality of the brain where the level of annoyance is contextually based.

The *objective information redundancy* is considered from an information communication perspective. The goal is for a transmitter of information to send the minimum amount of information to a receiver such that the desired information is exactly reconstructed. The 'minimum amount of information' directly implies redundancy in that it is not necessary to send all the information. For digital video data representations the amount of information is measured as the average entropy with a binary alphabet, A = {0, 1}, assuming a stationary random process model.

A *target channel bit rate* or, in the case of a buffered output, a *target short-term average bit rate*, implies that further information reduction may be required even after the redundancy has been removed. This is particular to low bit rate applications and the *optimal exploitation* requires the added distortion to result in a 'graceful degradation' of perceived quality. The degradation to meet the bit rate requirement is achieved by weighted information quantisation. The entire process is therefore a trade off between distortion and bit rate.

For a given video coding application, an arbitrary exploitation may be derived as a solution to the ill-posed problem statement. Is the derived exploitation optimal? For that matter, is any

exploitation a solution? The solution is considered to be valid as an algorithm if it meets the bit rate requirement but is bounded by subjective acceptance within the application context or in relation to another solution. Therefore the specifications of the video coding application are constructed, a solution is derived with the problem statement as the objective and the decoded video is evaluated for subjective acceptance. The engineering problem is to select an appropriate operating point on the distortion-rate function that is constructed for the particular application.

Consider the video telephony application at very low bit rates (≤ 10k bits/s). The audio bit stream has the highest priority and therefore the video must be able to 'keep up', at least perceptually. For ease of 'lip synchronisation' and 'live' operation there must be a minimum coding delay (≤ 100ms) which excludes the use of source coding channel buffers and INTRA start frames. To provide some reasonably smooth 'lip movement' with the audio, the frame rate should be at least 10 frames/s and should be constant. These constraints impose a constant number of bits per frame that must be able to be changed at one frame's notice to provide for channel bit rate fall-back capabilities. Note that the ITU-T Recommendation H.263 (H263 algorithm) designed for video telephony applications, is not well suited to these specifications.

A video coding solution with and without motion compensation is derived for this application. Having established the algorithm, discussions of the salient features that underpin the coding gain are necessary. The proposed algorithm is discussed followed by a detailed examination of the main quantisation functions that contribute to the compression gain. The algorithm is evaluated with a set of video test sequences and comparisons are made from a peak signal-to-noise ratio (PSNR) perspective.

2 Proposed Video Coding Algorithm

Each image in the video sequence is presented to the input of the algorithm at the required frame rate. The proposed video coding algorithm operates in a first order coding loop where only the instantaneous first derivative of the information is coded.

$$f'(z) = f(z)\left(1 - z^{-1}\right)$$

The efficiency of the loop relies on the temporal redundancy between images in a sequence. Note that higher frame rates have a greater temporal redundancy than lower rates for the same change of viewing scene. This suggests that at very low frame rates (< 2 frames/s) a coding

loop of this kind may not be the most efficient approach. The algorithm may be operated in a basic form and an extended form that includes motion compensation.

2.1 Basic Algorithm

A block diagram of the basic algorithm without motion compensation is shown in Figure 2.

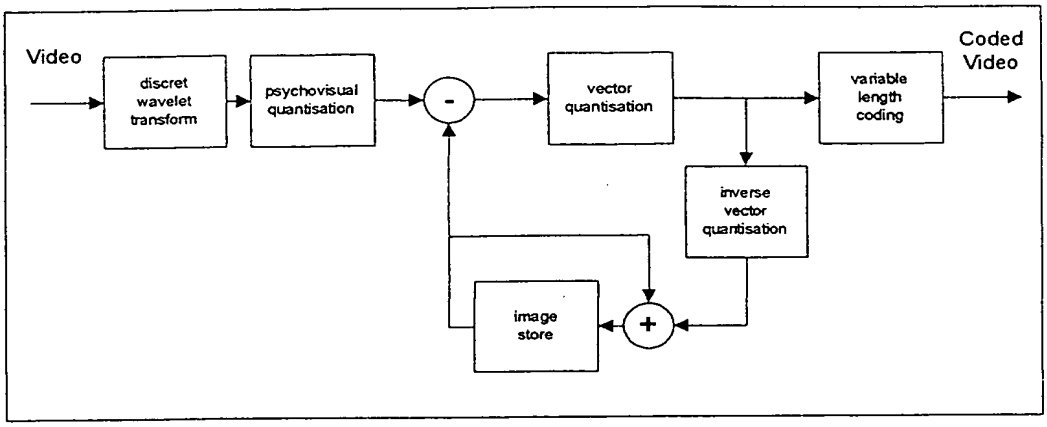


Figure 2: Block Diagram of Basic Video Coding Algorithm

For any scheme of this nature, maximum objective information exploitation where the output is entropy coded is achieved when the result of the information difference calculation produces small similar valued numbers. From a practical perspective this means mostly zeros and, if not zero as close to zero as possible. Therefore the coding efficiency is focussed on the input to the vector quantiser. The difference calculation is 'encouraged' to produce small values by firstly, stripping as much spatially redundant information from the input image as possible and then secondly, to subtract from it a best possible prediction of that input.

The fundamental underlying structure of this algorithm attempts to process the video data in a manner similar to that of the human visual system. In this way the algorithm is conducive to maximising the exploitation of the subjective information redundancy. From the simple model of the human visual system in Figure 1, the first area of redundancy exists in the eye's sensor mechanism. The eye is conditioned to be more sensitive to luminance information than chrominance. Therefore the video input data is represented by separate luminance (Y) and chrominance (U and V) components where the chrominance spatial resolution is reduced by two. The nomenclature for this colour space will be represented by the component's ratio as YUV 4:1:1. Colour separation and resolution reduction is a widely accepted front-end image compression method.

The visual cortex is responsible for the non-conscious processing of the visual information presented to it by the eye's sensor mechanism. The visual information is separated into textures and edges of various orientations and processed at differing resolutions. Therefore the multiresolution wavelet transform is chosen as an appropriate domain to operate within to simulate the way in which non-conscious processing is performed. The sensitivity to the separated sub-bands is not constant and a model of this function is used to scalar quantise the wavelet coefficients.

The wavelet transform is implemented as a discrete wavelet transform (DWT) using the lifting technique where the coefficients are derived from the 9-7 biorthogonal filter coefficients found to be suitable for image coding applications (Villasenor, Belzer and Liao 1995). The lifting technique provides greater advantages over the standard convolution method of implementing the DWT other than faster algorithms (Sweldens 1996). There is no distortion at the sub-band boundaries allowing the maximum number of resolution levels to be achieved such that the lowest resolution pixel dimensions may be of the same order as the filter coefficient support. From an objective information viewpoint the choice of the particular biorthogonal filter coefficients is crucial to producing image decompositions with the minimum number of significant valued coefficients. Naturally occurring continuous-tone images are mostly composed of large smooth areas and sharp edged boundaries (probably the reason for the evolutionary development of the nature of the human visual cortex). The visual data, particularly at edges, is not well suited to fixed resolution harmonic function decompositions in that they produce many significant valued coefficients. Therefore an appropriately chosen DWT will, in general, perform better than harmonic transforms, such as the discrete cosine transform, in a video coding environment.

The input image with the modelled subjective information removed has the stored predicted reference DWT image subtracted from it on a coefficient by coefficient basis. A set of sub-band vector quantisers, one for each level, orientation and colour component, is used to quantise the difference coefficient image. The bit rate is partially controlled by an algorithm parameter that thresholds the vector coefficients before quantisation. The codebook indices are entropy coded and therefore the output bit rate is dependent on the operational parameters of the vector quantiser. The operationally optimal vector quantiser for each sub-band is found by selecting the most suitable operating point on the distortion-rate function for that sub-band. Suitability is determined from both sub-space coverage and a practical entropy code perspective.

Entropy coded codebook indices generate variable length output bit streams that are dependent on the amount of change in information between images in the video sequence. One of the desired properties of the algorithm for video telephony applications is a constant bit rate per frame regardless of the information content. Therefore the coded bits must be allocated from the most important to the least important coefficients from a subjective viewpoint. The proposed algorithm incorporates two cognitive factors into the coding process. Firstly, objects are recognised as a whole before filling in the detail. In video telephony, a human head and shoulders are noted before attempting to recognise the individual. The low resolution comes before the high and therefore will be considered as more important for the algorithm. Secondly, human vision tends to be foveated especially while tracking a moving object. For video telephony, it is most likely that the face will be focussed on and will usually be in the centre of the image. Therefore it is most profitable for the coding to be centre biased. The algorithm combines these factors by coding each difference frame vector from the lowest resolution DWT sub-band to the highest and spirally from the centre outwards within the sub-band. The luminance colour components are favoured over the chrominance. The process is diagrammatically illustrated in Figure 3.

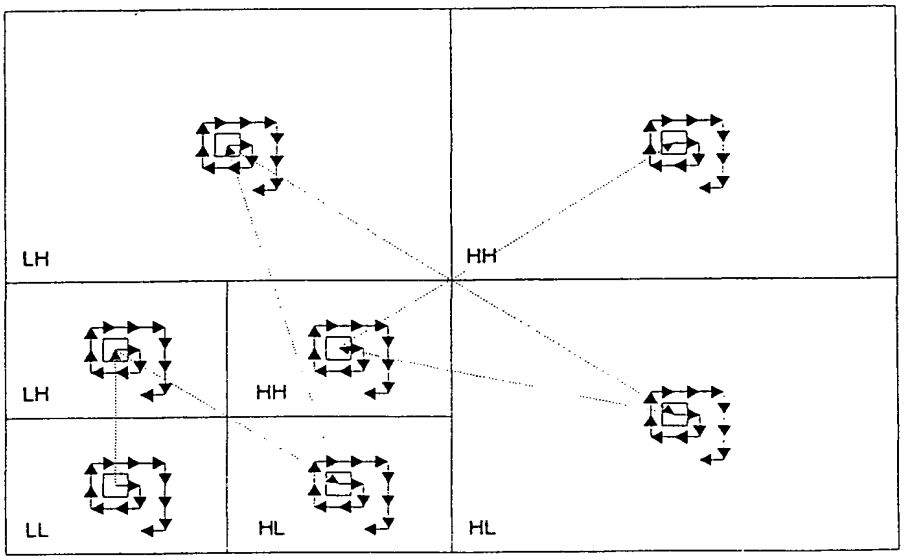


Figure 3: Foveated Coding Order for a Two Level DWT Decomposition

The majority of vectors contain all zero values therefore the coding proceeds in the ordering given by counting the zero vectors until a non-zero vector is encountered. The run of zero vectors and the codebook index of the non-zero vector are entropy coded with their respective codes. If the number of bits generated from this pair added to the current running total of bits is less than the frame bit limit then, the two entropy codes are added to the bit stream. This

process continues until the frame bit limit is reached and no further bits are transmitted to the receiver. All remaining vectors are assumed to be zero by the decoder. 'End-of-sub-band' and 'end-of-image' markers are included for bit stream synchronisation. Note that the 'end-of-sub-band' markers provide for the receiver to partially decode the bit stream such that a 'thumbnail' representation (lower resolution) may be viewed during reception. This is particularly useful for monitoring communications sessions or scanning stored video databases.

The artefacts generated from this coding method are a result of producing a variable length of coded vectors per frame in the ordered sequence. During periods of low temporal activity the ordered vector sequence will reach the high resolution sub-bands for the given frame bits. If this is followed by a burst of temporal activity then, the ordered sequence will only reach lower resolution sub-bands for the same frame bits and implicitly code zero change in the previously reached higher sub-bands. The visual effect is to leave motionless high frequency edge components superimposed on the moving image. This artefact is referred to as the 'shower door' effect. To alleviate the effect, the remaining higher resolution sub-band vectors in the decoded image are set to zero from the centre out until the same foveated point reached by the last coded sub-band. To allow for differing vector dimensions in each sub-band the foveated point is calculated as a fraction of the sub-band ordered sequence.

2.2 Motion Compensation

To further reduce the energy of the information presented to the vector quantiser, it is possible to track spatial translational motion between images in the video sequence. Current international standards, for example, the H263 algorithm, achieve most of their compression gain from the motion compensation process in the spatial domain. It is possible to apply similar block motion estimation techniques in the DWT domain but the performance of the approximation is limited. The limitations are discussed below in Section 5. DWT domain motion compensation has some advantages over the spatial domain counterpart in terms of the visual artefacts generated in those cases where the approximation error is large. Spatial domain compensation produces annoying blocking effects, whereas the speckled noise produced from the DWT domain compensation is less objectionable.

The basic algorithm is extended to include DWT domain motion estimation and compensation. The complete block diagram of the algorithm is shown in Figure 4.

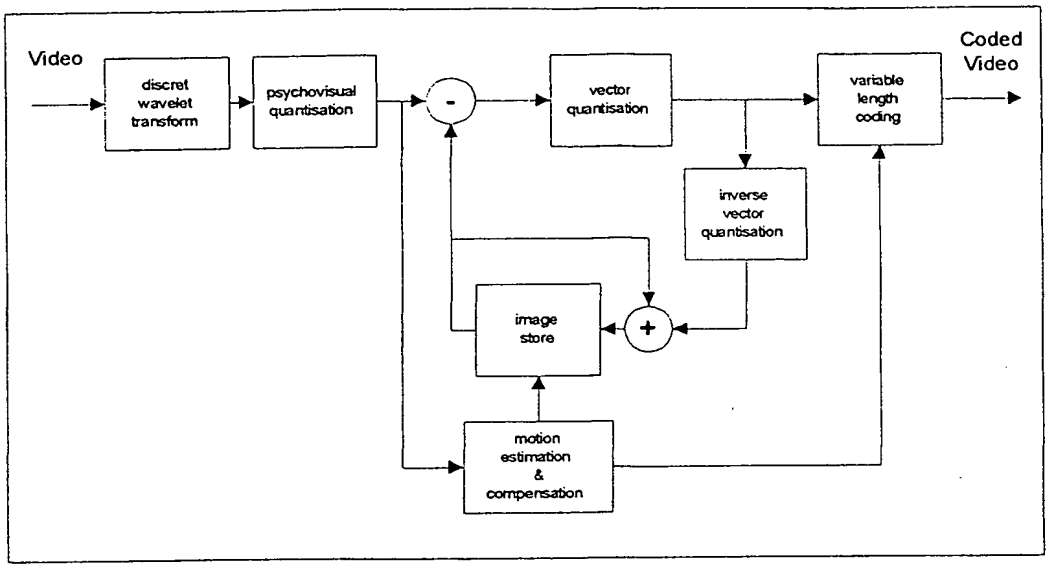


Figure 4: Block Diagram of DWT Domain Video Coding Algorithm

The basic principle of the DWT domain motion estimation and compensation proceeds in the following manner. After the subjective information has been removed from the current image, motion is estimated with each two-dimensional block of spatially corresponding coefficients within each sub-band and colour component of the stored reference image. Within each sub-band the block dimension is chosen to be the same as that of the corresponding vector quantiser, to allow the foveated sequence to proceed on a block-by-block basis without overlap. Extending the sub-band boundaries with zero valued coefficients permits motion vectors from outside the sub-band. This increases the estimation accuracy for the boundary blocks. Half-pixel estimation is performed and the best MSE matching block is written into the reference image. The compensated reference image is then subtracted from the input image and vector quantised.

There is generally little motion between consecutive frames in a video sequence particularly in the background. Therefore zero vectors will be statistically most likely. The motion vector foveated sequence is coded in a similar manner to the run length encoding of zero vectors applied to the indices of the vector quantisers. For each non-zero motion vector in the sequence, a zero run length and the two motion vector components are entropy coded and added to the output bit stream provided the frame bit limit is not exceeded.

The zero run length encoding of the motion vectors is particularly important to very low bit rate video coding algorithms where the added bit overhead may offset the quality gained by the process. For small movement within video scenes, as possible in video telephony, the

many zero and small valued motion vectors will consume scarce bit resources. An added problem with block motion estimation is that it is possible for the compensated block to produce greater difference image energy than without compensation. An uncompensated reference image block is equivalent to a compensated block with a zero motion vector. Therefore to improve the predicted reference image and to 'encourage' the zero motion vector for coding efficiency, a 'best choice' algorithm is used.

The 'best choice' algorithm is in a block MSE sense. The basis of the choice is determined by vector quantising the difference image blocks from both the compensated and uncompensated images and choosing that with the lowest quantisation error. If the uncompensated image block is chosen then, the zero vector is associated with it. The process is diagrammatically illustrated in Figure 5.

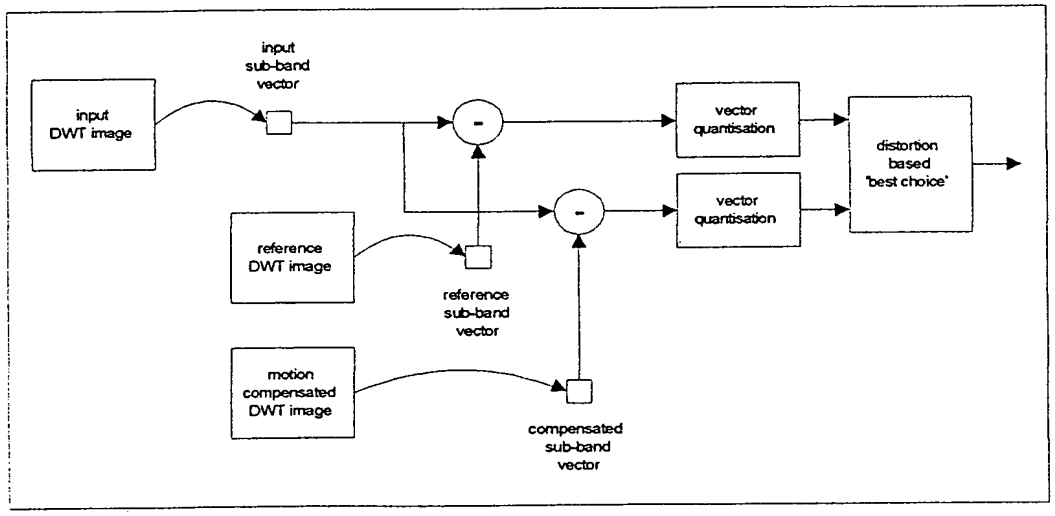


Figure 5: Block Diagram of the 'Best Choice' Algorithm

The decoder does not require any knowledge of the choice since the zero vector implicitly refers to an uncompensated block.

The extended form of the algorithm that includes motion compensation has large performance gains during periods of large temporal movement. But even in very low bit rate environments with low temporal activity, the 'best choice' algorithm with run length coding of the zero motion vectors ensures a minimal bit cost.

3 Psychovisual Shaping of DWT Coefficients

The degree to which spatial frequency representations of images may be shaped depends on the frequencies observed by a human viewer. Therefore an expression for the relationship between spatial frequency and the geometry of the viewing environment is required. The horizontal and vertical dimensions are treated similarly but independently. Consider the horizontal dimension of a display of width, w meters, viewed from a distance of h meters shown in Figure 6. The subtended viewing angle, a degrees, is given as follows:

$$a = 2\tan^{-1}\left(\frac{w}{2h}\right) \quad [\text{deg}]$$

The maximum frequency that may be represented by a digital display is the Niquist rate of half the horizontal pixel rate, r. Therefore the maximum spatial frequency, f_{max}, is given by:

$$f_{max} = \frac{r}{4\tan^{-1}\left(\frac{w}{2h}\right)} \quad [\text{cycles/deg}]$$

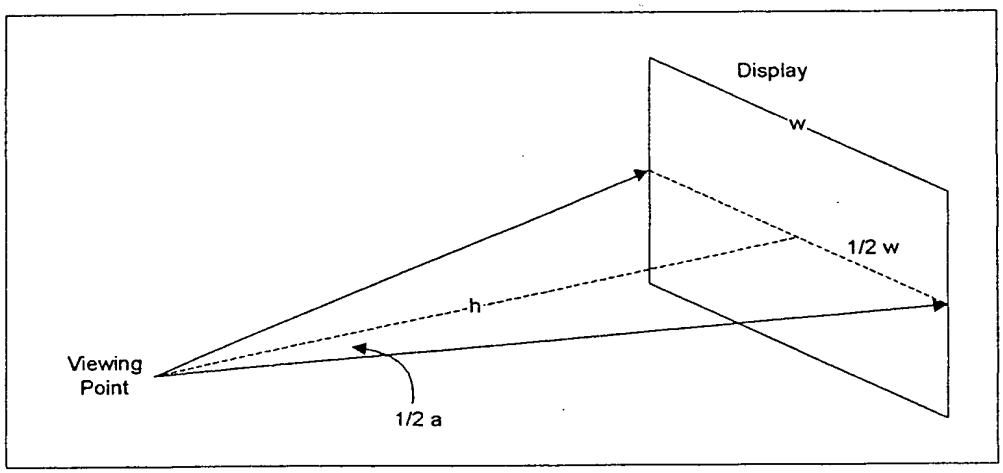


Figure 6: Display Viewing Environment

The spatial frequency response of the human psychovisual system is known to have a band-pass shape. The response has been found by subjective experiments involving a 'just-noticeable-threshold' approach to produce a standard contrast sensitivity function (Sakrison 1977). One of the more common fitted models for the function verses spatial frequency, f, (Ngan, Leong and Singh 1989) is defined as follows:

$$S = (0.31 + 0.69f)e^{-0.29f}$$

The function is plotted in Figure 7.

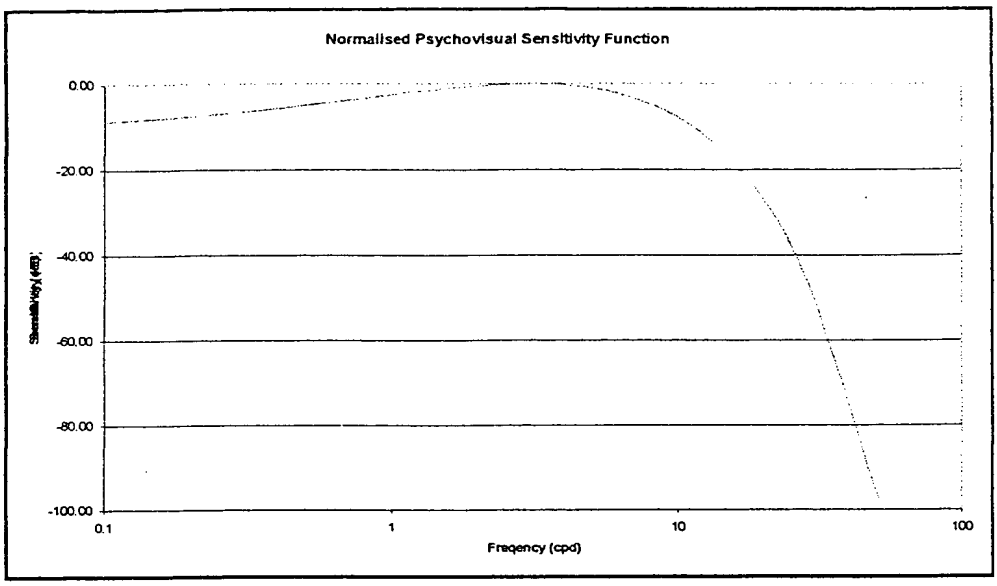


Figure 7: Psychovisual Sensitivity Function

This model has been successfully applied to constructing quantisation tables for 8x8 discrete cosine transform (DCT) based luminance-only image compression (McLaren and Nguyen 1991). This model provides for an overall frequency response from a high level perspective only. It does not give any indication of what low level visual cortex processing is taking place to achieve this response.

There is some experimental evidence to show that the retinal image is decomposed into frequency bands (Campbell and Robson 1968) and processed with a varying sensitivity to visual orientation (Campbell and Kulikowski 1966). The visual filters have an average bandwidth of approximately 1.3 octaves with the highest sensitivity at 0° and 90° visual angle and least sensitive at 45°. This knowledge has been used to develop a perceptual quality metric to include the human psychovisual and physiological aspects for evaluating image and video coding algorithms (van den Branden Lambrecht et al 1996). The results may, at least, be construed as experimental evidence that the model is valid.

The two-dimensional DWT consists of a frequency and orientation sub-division process. The multiresolution filtering and sub-sampling in the horizontal and vertical directions divides the signal into octave frequency sub-bands with horizontal, vertical and diagonal orientations. The DWT process may therefore be considered as a discrete approximation to the physiological process that takes place in the model for the human visual cortex (Mallat 1989).

An efficient quantisation strategy may be developed to take advantage of this model. The visibility of quantisation errors for all colour components at all DWT levels and orientations has been determined from a set of subjective experiments (Watson, Yang, Solomon and Villasenor 1996). In this way, a psychovisual model customised for the two-dimensional DWT coefficients is established. The resulting fitted model for the threshold of visibility, T, as a function of spatial frequency, f, and orientation, o, is as follows:

$$T(f,o) = a.10^{k\left[\log\frac{f}{g_o f_0}\right]^2}$$

The approximated model parameters are given in Table 1.

Table 1: Threshold Model Parameters

Colour	a	k	f_0	g_{LxLy}	g_{LxHy}	g_{HxLy}	g_{HxHy}
Y	0.495	0.466	0.401	1.501	1.0	1.0	0.534
U	1.633	0.353	0.209	1.520	1.0	1.0	0.502
V	0.944	0.521	0.404	1.868	1.0	1.0	0.516

The successive application of the DWT at each level, l, results in the halving of the sub-band frequency bandwidth. For a display with a maximum spatial frequency of f_{max}:

$BW_l = f_{max} 2^{-l}$ [cpd]

The centre frequency of each sub-band is used as the nominal frequency for the development of the quantisation process.

$f_{l,c} = 3f_{max} 2^{-(l+1)}$ [cpd]

The visibility of quantisation errors introduced at a particular level and orientation may be approximated by the amplitude of the DWT synthesis filters for that level and orientation. This approximation is implementation and filter bank dependent. Consider the introduction of quantisation errors at level (m+1) for the inverse DWT process shown in Figure 8 (Antonini, Barlaud, Mathieu and Debechies 1992).

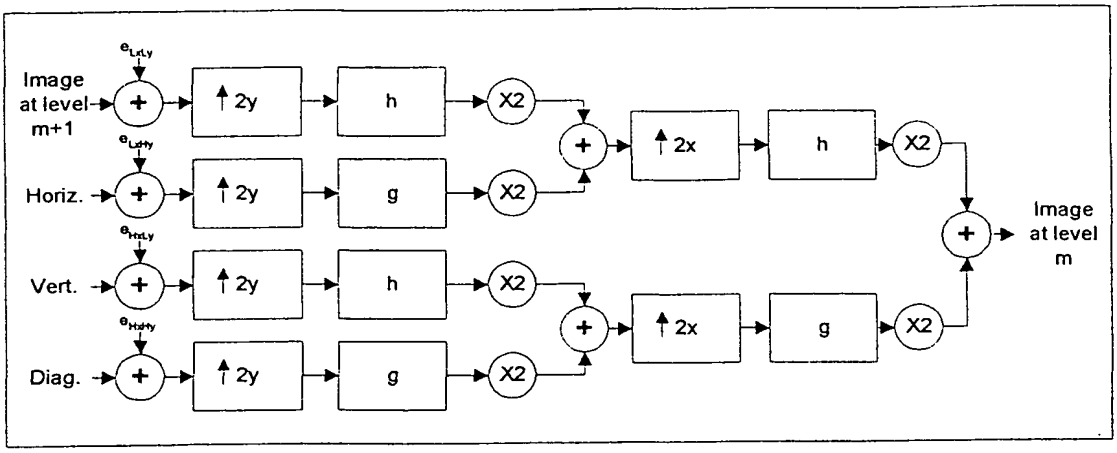


Figure 8: Two-dimensional Inverse DWT for a Single Level

Each sub-band is up-sampled by two in the vertical direction, then a convolution is performed with the relevant one-dimensional synthesis filter, multiplied by a factor of two and summed to form vertical groups. The vertical groups are similarly up-sampled followed by the convolution, multiplication by two and summed, but in the horizontal direction The resulting effect on the image at level m is then propagated through the remaining image levels to the reconstructed image. Considering only the most significant term of the linear convolution process approximates the amplitude of the error per level. For example, the effect of an error in a *low-pass horizontal and a high-pass vertical* orientation (LxHy) at one level may be approximated as follows, for the next level.

$$E_{LxHy} = \left| \left(e_{LxHy} * g * h\right) 2^2 \right| \cup 2^2 \left| e_{LxHy}\, g_0\, h_0 \right|$$

Therefore the effect of an error at level m for each orientation, on the entire reconstructed image may be written as:

$$E_{m,LxLy} \cup 2^{2m} \left| e_{m,LxLy}\, h_0^{2m} \right|,$$

$$E_{m,LxHy} \cup 2^{2m} \left| e_{m,LxHy}\, g_0 h_0 h_0^{2(m-1)} \right|,$$

$$E_{m,HxLy} \cup 2^{2m} \left| e_{m,HxLy}\, h_0 g_0 h_0^{2(m-1)} \right|,$$

and,

$$E_{m,HxHy} \cup 2^{2m} \left| e_{m,HxHy}\, g_0^2 h_0^{2(m-1)} \right|.$$

A biorthogonal wavelet filter bank that performs well for image compression applications, is the spline-based set with filter coefficient lengths of nine and seven (Villasenor et al 1995). The approximate amplitudes of error visibility for this filter bank with a root-two bias as

required for the inverse DWT described above, to four levels and all orientations of the DWT process are given in Table 2.

Table 2: Quantisation Error Visibility to DWT Level 4

	Level			
Orientation	**1**	**2**	**3**	**4**
LxLy	1.2430	1.5461	1.9224	2.3904
LxHy	1.3447	1.6720	2.0790	2.5851
HxLy	1.3447	1.6720	2.0790	2.5851
HxHy	1.4542	1.8082	2.2483	2.7956

A quantisation factor is required for each colour and sub-band such that the resulting quantisation error is below the visibility threshold. For a linear quantiser with a factor of Q, the worst case error is Q/2. Therefore the quantisation strategy is defined as:

$$T(f,o) = V(l,o)\,\frac{Q(l,o)}{2}$$

The quantisation visibility term, V is defined by the DWT process such as that given in Table 2. The operational quantisation factors are formed as follows:

$$Q(l,o) = \frac{2}{V(l,o)}\,T(f,o)$$

$$Q(l,o) = \frac{2}{V(l,o)}\,a.10^{k\left[\log\frac{f_{i,c}}{g_o f_0}\right]^2}$$

$$Q(l,o) = \frac{2}{V(l,o)}\,a.10^{k\left[\log\frac{3f_{max}}{2^{(i+1)}g_o f_0}\right]^2}$$

These quantisation factors provide an overall shape that may be applied to the DWT coefficients of an image to achieve an imperceptible difference with the original image. For low bit rate applications where greater quality loss is tolerated, the quantisation shape is uniformly scaled to ensure that the largest errors are confined to the least responsive regions of the human psychovisual system.

4 Vector Quantisation with SOM's

The approach to generate a vector quantiser solution requires formulating a cost function to be minimised and describing the conditions of optimality for that cost function. The general vector quantisation theory is discussed to lay a foundation for showing how the trained SOM achieves the same optimal solution. The noise model of the vector quantiser produces a similar gradient descent training algorithm as that of the SOM. Operationally optimal vector quantisers for a signal compression environment may be found by varying the entropy of the SOM. The entropy is approximated as a rate-constraint and results in a natural extension to the SOM training algorithm.

4.1 Basic Vector Quantisation

Consider the basic encoder-decoder model for a vector quantiser as shown in Figure 9 (Gersho and Gray 1992). The processes discussed all assume the high resolution case where the number of codewords, N, is very large i.e. $N \rightarrow \infty$.

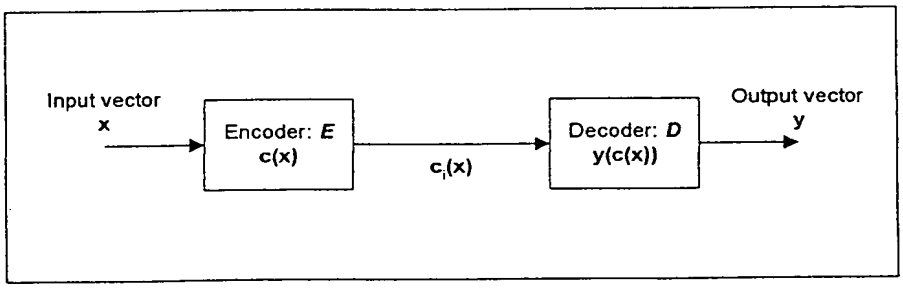


Figure 9: Basic Encoder-Decoder Vector Quantisation Model

The vector quantisation problem may be considered as an optimal approximation process of an input space to an output space that, in the case of signal compression, is itself a subset of the input space. A vector, $\mathbf{x}$, of dimension k from the input space, $\mathbf{X}^k$, (where $\mathbf{X} \in \mathbf{R}$, the set of real numbers) is the input to the encoder. The encoder maps the input space to a codebook $\mathbf{C}$ that consists of N codewords. The encoding process may be written as: E: $\mathbf{X}^k \rightarrow \mathbf{C}$ where $\mathbf{C}=\{c_i(\mathbf{x})\}_{i \in I}$. The process is fully defined by either the index into the codebook, $i \in \mathbf{I}$, or the codeword itself, $c_i(\mathbf{x})$, and therefore it is usual for only the index, or a symbolic representation thereof, to be transmitted on the communications channel to the decoder. The encoding process is the quantisation process that, in general, is lossy in that the codebook size is limited, $|\mathbf{C}| = \max\{i\} = N << \infty$. The decoder maps the codebook back into the input space and may be written as: D: $\mathbf{C} \rightarrow \mathbf{X}^k$ where the reconstructed vector $y \in \mathbf{X}^k$.

The compression mechanism of vector quantisation is achieved by the dimension reduction process from a vector in space $\mathbf{X}^k$ to an integer index, i. The premises for the mechanism is that the signal space covered by $\mathbf{x}$ is a sub-space of $\mathbf{X}^k$ and that it is a stationary random process with an underlying joint probability density function (pdf), $f(\mathbf{x})$, such that $f(\mathbf{x})\rightarrow 0$ as $\mathbf{x}\rightarrow\pm\infty$ defined as $\{x_1\rightarrow\pm\infty\ ,\ x_2\rightarrow\pm\infty\ ,\ldots,\ x_k\rightarrow\pm\infty\}$.

Generating optimal or near optimal vector quantisers for the signal sub-space is achieved by minimising a suitable cost function in a long term average sense where the cost function is itself defined as a stationary random process. If the cost function and the sub-space pdf are smooth functions, or may be approximated by smooth functions, and hence are differentiable everywhere, then gradient descent methods may be used to find near optimal solutions for sometimes intractable analytical solutions. The most common approach is to minimise the mean squared error (MSE) distortion in a Eucledian distance sense, because the performance of the resulting vector quantiser is usually measured by the MSE criterion. The function to minimise may be written as:

$$D = \int_{-\infty}^{+\infty} \|\mathbf{x}-\mathbf{y}\|^2 f(\mathbf{x})d\mathbf{x} \qquad \mathbf{x},\mathbf{y} \in R^k$$

Here the $\|\bullet\|$ operator represents the Eucledian distance. The optimal solution to the minimisation of D with respect to $\mathbf{y}$, requires the joint minimisation of the *nearest-neighbour* condition and the *centroid* condition.

The *nearest-neighbour* condition describes the optimal encoder given a fixed decoder. This condition results in the input space being partitioned into regions, $\mathbf{R}_i$, which may be termed as k-dimensional "volumes of influence" of the fixed codewords, $\mathbf{c}_i = \mathbf{y}_i$, in the codebook, $\mathbf{C}$. The optimal region partitions are such that:

$$\mathbf{R}_i \in \{\mathbf{x} : \|\mathbf{x}-\mathbf{c}_i\|^2 \le \|\mathbf{x}-\mathbf{c}_j\|^2\} \qquad j = 1\ldots N$$

Therefore:

$$\|\mathbf{x}-\mathbf{y}\|^2 = \min_{\mathbf{c}_i}\{\|\mathbf{x}-\mathbf{c}_i\|^2\}$$

The region is chosen to minimise the squared error distortion with the given codebook.

The *centroid* condition describes the optimal decoder given a fixed encoder. The distortion integral, D, may be rewritten as:

$$D = \sum_{i=1}^{N} \int_{R_i} \|x - c_i\|^2 f(x)dx$$

$$= \sum_{i=1}^{N} P_i \int \|x - c_i\|^2 f(x \mid x \in R_i)dx$$

Here P_i is the probability that $\mathbf{x}$ is in $\mathbf{R}_i$, $P_i = \text{Prob}[x \in \mathbf{R}_i]$ and $f(\mathbf{x}|\mathbf{x} \in \mathbf{R}_i)$ is the conditional pdf of $f(\mathbf{x})$ given that $\mathbf{x}$ lies in $\mathbf{R}_i$. The fixed encoder implies that the regions, $\mathbf{R}_i$, are fixed and therefore each conditional term may be separately minimised, provided that P_i is non-zero. Therefore the centroid of region, $\mathbf{R}_i$, is defined as that output vector, $\mathbf{y}_i = \mathbf{c}_i$, which minimises the distortion between itself and the input vector, $\mathbf{x}$, where $\mathbf{x} \in \mathbf{R}_i$, over the entire conditional pdf.

$$y_i = \min_{y}^{-1} \left[\int_{R_i} \|x - y\|^2 f(x \mid x \in R_i)\, dx \right]$$

Under the squared error distortion criterion the optimal solution is the centroid of each region.

$$y_i = \int_{R_i} x\, f(x \mid x \in R_i)dx$$

An iterative batch mode algorithm may be used to find an optimal or near-optimal solution to these two conditions for a given input distribution. The process involves finding an optimal region partition for a given codebook and then finding the optimal codebook for the given partitions. Many such algorithms and operational derivatives exist for this process (Linde, Buzo and Gray 1980, Gresho et. al. 1992).

4.2 Vector Quantiser Noise Model

Consider the gradient descent training process of an optimal high resolution vector quantiser. During the early stages of the training process there is a large error between the input, $\mathbf{x}$, and output, $\mathbf{y}$. As the process continues and the global minimum is approached, the error decays to some small value, which, in the high resolution case, may be made arbitrarily small. If the error is assumed to be independent of the input, then it may be modelled by a smooth zero-mean Gaussian distributed random variable. The model in Figure 9 may be modified to produce the noise model of a vector quantiser (Luttrell 1989, Haykin 1994) shown in Figure 10.

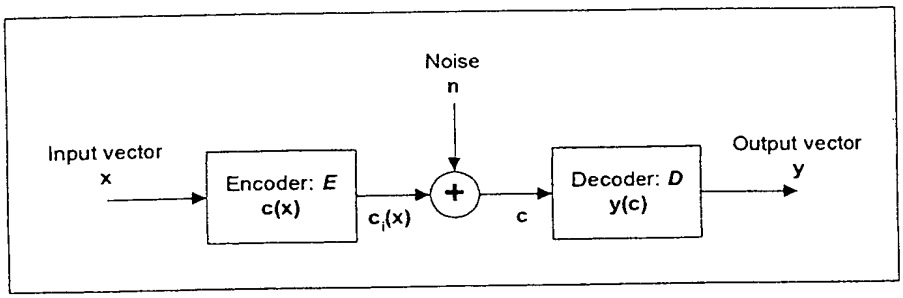


Figure 10: Encoder-Decoder Noise Model

For this model, the optimal vector quantiser for vectors, $\mathbf{x}$, taken from a sample space defined by the underlying pdf, $f(\mathbf{x})$, in a squared error sense, is one that minimises the long term average distortion defined as:

$$D = \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{+\infty} \|\mathbf{x} - \mathbf{y}(\mathbf{c}(\mathbf{x}) + \mathbf{n})\|^2 \, \pi(\mathbf{n}) \, d\mathbf{n} \right] f(\mathbf{x}) \, d\mathbf{x}$$

Here $\pi(\mathbf{n})$ is the pdf of the additive noise. The optimal encoder for a given decoder then minimises the partial distortion measure:

$$pD_{c,x} = \int_{-\infty}^{+\infty} \|\mathbf{x} - \mathbf{y}(\mathbf{c}(\mathbf{x}) + \mathbf{n})\|^2 \, \pi(\mathbf{n}) \, d\mathbf{n}$$

For a given input vector, $\mathbf{x}$, find the minimum distortion codeword, $\mathbf{c}(\mathbf{x})$, for all possible noise additions. The realisation of this equation may be simplified by assuming that $\pi(\mathbf{n})$ is smooth and the training process is nearing completion, so $\mathbf{n} \rightarrow 0$ and therefore $\pi(\mathbf{n}) \rightarrow 1$. The minimum condition then reduces to the *nearest-neighbour* condition of the previous noiseless model. The best-matching codeword, $\mathbf{c}(\mathbf{x})$, from the vector quantiser codebook, $\mathbf{C}$, may be completely defined by its index, $i \in \{1...N\}$, in the codebook (and vise versa).

$$i(\mathbf{x}) = \min_j{}^{-1} \left\{ \|\mathbf{x} - \mathbf{y}(\mathbf{c}_j(\mathbf{x}) + \mathbf{n})\|^2 \right\} \quad j = 1...N$$

The optimal decoder for a given encoder is found by minimising the distortion measure with respect to the output, $\mathbf{y}$.

$$\frac{fD}{fy} = -2 \int_{-\infty}^{+\infty} [(\mathbf{x} - \mathbf{y}) \pi(\mathbf{n})] f(\mathbf{x}) \, d\mathbf{x}$$

Setting to zero and solving results in the *centroid* condition that may be used in an iterative batch mode algorithm.

$$\mathbf{y} = \frac{\int_{-\infty}^{+\infty} \mathbf{x}\,\pi(\mathbf{n})\,f(\mathbf{x})\,d\mathbf{x}}{\int_{-\infty}^{+\infty} \pi(\mathbf{n})\,f(\mathbf{x})\,d\mathbf{x}}$$

However, a gradient descent algorithm for the decoder output vectors follows directly. Note that $\mathbf{y}$ is a function of $\mathbf{c}$, and $\mathbf{n} = \mathbf{c} - \mathbf{c}(\mathbf{x})$. Therefore, randomly sampling the input space that has a distribution defined by the pdf, $f(\mathbf{x})$, results in the following update step:

$$\mathbf{y}_{t+1}(\mathbf{c}_j) = \mathbf{y}_t(\mathbf{c}_j) + \eta_t \pi_t(\mathbf{c}_j - \mathbf{c}_i(\mathbf{x}_t))[\mathbf{x}_t - \mathbf{y}_t(\mathbf{c}_j)], \quad j = 1...N$$

Here η_t is the adaptive learning rate parameter and $\mathbf{c}_i(\mathbf{x}_t)$ is defined by the best-matching codeword from the *nearest-neighbour* condition at iteration t. If a zero-mean Gaussian function for $\pi(\mathbf{n})$ is imposed on the learning process, then the codewords, $\mathbf{c}_j$, centred on the best-matching codeword, $\mathbf{c}_i(\mathbf{x}_t)$, will be 'brought nearer' in a squared error sense, to the centre codeword. Furthermore, by imposing an initially large radius Gaussian function and then decreasing it to zero during the training process, will result in a topologically ordered codebook, $\mathbf{C}$.

The stochastic gradient descent algorithm for the encoder-decoder noise model is the same as the standard SOM algorithm with neuron indices defined by $i = \{1...N\}$ and the neuron weights defined by $\mathbf{y}_i$. The noise shape defines the neighbourhood function and ensures the topologically ordered property of the output space. Therefore the SOM algorithm will generate an optimal (or at least, near optimal) vector quantiser.

4.3 Rate-Constrained SOM

Consider now the application of a SOM trained as an optimal vector quantiser in a signal compression environment. The vector samples, $\mathbf{x}$, are extracted from the signal and the index of the neuron whose weight vector has the lowest squared error distortion with $\mathbf{x}$, is transmitted on the channel to the receiver. The receiver decodes the output vector, $\mathbf{y}$, in a weight vector look-up table whose neuron index, i, is that received from the channel. Therefore the information conveyed over the communication channel is completely contained in the index. This raises the issue of efficient symbolic representation of the transmitted indices. Since only binary alphabet symbols are considered here, $A \in \{0, 1\}$, the index is represented by a variable length code, v(i), whose average bit rate is upper-bounded by the uniform neuron firing distribution case of $B = \log_2 N$ bits per vector. The bit rate or bit length of the code will be denoted by its magnitude, $|v(i)|$.

For any joint pdf $f(\mathbf{x})$ of the input vector, $\mathbf{x}$ with dimension k, such that $f(\mathbf{x})\rightarrow 0$ as $\mathbf{x}\rightarrow\pm\infty$, there exists an arbitrary low distortion SOM where N is finite and the neuron firing probabilities are described by a probability mass function (pmf), $p(i(\mathbf{x}))$, such that $p(i(\mathbf{x})) \propto f(\mathbf{x})$. This premise is based on the density matching properties of the trained SOM (Kohonen 1997, Haykin 1994). An entropy coding method is therefore more efficient for transmitting the indices. From an information viewpoint the average entropy, in bits, is defined as:

$$H(i) = -\sum_{i=1}^{N} P_i \log_2 P_i$$

Here P_i is the *a posteriori* probability of index, i being the winning neuron which is identical to the vector quantiser definition, $P_i = \text{Prob}[\mathbf{x} \in \mathbf{R}_i]$. If a prefix-free variable length binary code is used and allowing non-integer code lengths, then the average length is the index entropy (Gresho et. al. 1992). Therefore the length of the code to represent the index, i, is defined as:

$$|v(i)| = -\log_2 P_i$$

Note that practical entropy codes must posses integer multiple lengths and therefore this acts as an asymptotic lower bound. The long term average bit rate generated from the trained SOM is written as:

$$B = \sum_{i=1}^{N} P_i |v(i)|$$

The rate-constraint is constructed from the trade-off between rate, (required for compression), and distortion (required for image quality). For low bit rate coding we wish to sacrifice image quality for higher compression, but a quantitative analysis in the given coding environment is required in order to make a prudent choice of an operational point. An approximate operational distortion-rate function, for a fixed vector dimension and a given vector quantiser, $(\mathbf{c}(\mathbf{x}), \mathbf{y}(\mathbf{c}))$, may be constructed as a lower bound from the distortion-rate theory underlying entropy-constrained vector quantisation (Chou, Lookabaugh and Gray 1983).

$$D^*(R) = \inf_{(\mathbf{c}(\mathbf{x}),\mathbf{y}(\mathbf{c}))} \{ E[\,\|\mathbf{x}-\mathbf{y}\|^2\,] \mid E[R] \le \log_2 N \}$$

Here E[ ] is the statistical expected operator of a stationary random process. This equation defines the bounds of the region in the distortion-rate plane within which we are able to operate with the parameter restrictions of our particular coding environment.

Consider the trained SOM where the neighbourhood function has decayed to zero and hence, assuming the high resolution case, the noise pdf, $\pi(\mathbf{n})$ is approximated by the Dirac function. The operational average distortion, D^* is a minimum in the sense that it minimises the function:

$$D = E[\,\|\mathbf{x}-\mathbf{y}\|^2\,] = \int_{-\infty}^{+\infty} \|\mathbf{x}-\mathbf{y}\|^2 f(\mathbf{x})\,d\mathbf{x}$$

The *nearest-neighbor* and *centroid* conditions result in an optimal partitioning of the SOM weight space into the regions, $\mathbf{R}_i$. The rate constraint, B which is approximated by H(i) and introduced with a Lagrange multiplier, λ, to generate the operational distortion-rate function to be minimised.

$$D^*(R) = \sum_{i=1}^{N} P_i \int_{-\infty}^{+\infty} \left(\|\mathbf{x}-\mathbf{y}_i\|^2 - \lambda \log_2 P_i \right) f(\mathbf{x}\,|\,\mathbf{x} \in \mathbf{R}_i)\,d\mathbf{x}$$

The integral for the noise pdf, $\pi(\mathbf{n})$, has been omitted for simplicity of representation but is required for the training process as the neighbourhood function. Note that the $\lambda \log_2 P_i$ term is a constant and therefore affects the *nearest-neighbour* condition but does not affect the *centroid* condition. The rate-constrained SOM is trained with the gradient decent algorithm to minimise the instantaneous cost function described as:

$$J_\lambda = \|\mathbf{x}-\mathbf{y}_{i(\mathbf{x})}\|^2 - \lambda \log_2 P_{i(\mathbf{x})}$$

An approximation for the *a posteriori* $P_{i(\mathbf{x})}$ is made at each step of the descent algorithm.

4.4 Rate-Constrained SOM Algorithm

For the training process of the SOMs, that will be used as vector quantisers, to be considered meaningful, an appropriate training set must be formed. The training set must represent the input space in the sense that its pdf is an approximation to pdf of the input space, which will be applied to the vector quantiser during normal operation. In a video coding environment where the difference images are quantised within a first order coding loop, the quantisation error will be propagated to the reference image. The error will be superimposed on the next difference image and will appear as an input to the vector quantiser. In this way quantisation

errors will accumulate in time as subjectively annoying artefacts. This effect is exaggerated in low bit rate codecs where the number of neurons in the SOMs is restricted and the high resolution assumptions are not strictly valid. The problem may be reduced by dynamically adapting the training set during the training process.

The zero vectors are generally coded via some other mechanism (either a run-length approach or using a code/no code bit) and hence the training set consists only of non-zero difference vectors. The training set is formed from typical image differences and the addition of errors that the vector quantisers will make *after* training. However, knowledge of the final state of the quantisers is unknown during the training process, therefore error approximation is based on the current state. The applied training set is adapted at each iteration during the training process in the following manner. Two random samples, $\mathbf{x}_1$ and $\mathbf{x}_2$, are selected from the difference image training set, T, at iteration t.

$\{\mathbf{x}_1(t), \mathbf{x}_2(t)\} \subset T$

The update process of the SOM is performed with the first sample, $\mathbf{x}_1$. The error term, $\mathbf{e}_t$, for this sample and the winning neuron weight vector, $\mathbf{c}_1$, is determined. Simulating the effect of this error in the coding loop generates the second training sample, $\mathbf{x}_e$, for the SOM. Therefore the training data, T_t, applied to the SOM at iteration t, may be described as:

$T_t = \{\mathbf{x}_1(t), \mathbf{x}_e(t)\}$

Here;

$\mathbf{x}_e(t) = \mathbf{x}_2(t) - (\mathbf{x}_1(t) - \mathbf{c}_1(t))$

The training process continues in this way with a difference image sample and an error adapted sample at each iteration. After initialising the SOM neuron weights with small random values and setting the occurrence frequency, F_j to unity for each neuron, the training proceeds by repeating the steps 1 to 5 defined as follows:

Step 1: Randomly sample the training space to establish $\{\mathbf{x}_1(t), \mathbf{x}_2(t)\}$.

Step 2: Determine the first winning neuron, i_1:

$$i_1 = \min_j{}^{-1} \left\{ \left\| \mathbf{x}_1(t) - \mathbf{c}_j(t) \right\|^2 - \lambda \log_2 P_j \right\} \quad j = 1..N$$

$$P_j = \frac{F_j}{\prod_{j=1}^{N} F_k}$$

Step 3: Update the neighbourhood neuron weight values and the winning neuron occurrence frequency:

$$\mathbf{c}_j(t+1) = \mathbf{c}_j(t) + \eta(t)\pi_{j,i_1}(t)[\mathbf{x}_1(t) - \mathbf{c}_j(t)], \quad j = 1..N$$

$$F_{i_1} = F_{i_1} + 1$$

Here $\eta(t)$ is the exponential decaying learning rate parameter and $\pi_{j,i1}(t)$ is the Guassian exponential neighbourhood function with a linearly decaying radius centred on the winning neuron, i_1.

Step 4: Determine $\mathbf{x}_e(t)$ and find second winning neuron, i_2:

$$i_2 = \min_j{}^{-1} \left\{ \left\| \mathbf{x}_e(t) - \mathbf{c}_j(t) \right\|^2 - \lambda \log_2 P_j \right\} \quad j = 1..N$$

Step 5: Update the second winning neuron weight values and occurrence frequency:

$$\mathbf{c}_j(t+1) = \mathbf{c}_j(t) + \eta(t)\pi_{j,i_2}(t)[\mathbf{x}_e(t) - \mathbf{c}_j(t)], \quad j = 1..N$$

$$F_{i_2} = F_{i_2} + 1$$

4.5 Heuristics for Constraint Parameter Selection

For a DWT-domain video coding environment each DWT sub-band exhibits differing statistical properties and probability distributions therefore optimal rate-distortion selections are based on different criteria depending on the sub-band. The sub-band multidimensional probability distributions can not be assumed to be Gaussian nor are infinite resolution approximations necessarily applicable to finite size SOMs. The training set is a sample collection of the input space and therefore the underlying pdf is not smooth.

In practical implementations the choice of sub-band vector dimension is limited by the image size. For example, QCIF image sizes permit vector dimensions of 4x4, 2x2, 2x2 and 1x3 at sub-band levels of 1, 2, 3 and 4, respectively. Furthermore, the SOM dimensions are restricted by the need for practical variable length codes. SOM sizes of 8x8, 16x16, 32x32 and 64x64 neurons are considered practical. Operational rate-distortion points verses λ plots are

generated for all sub-bands and colour components and used to empirically select generic optimal points for constructing the vector quantisers.

Consider the results for the luminance, level 3 and LxHy DWT sub-band at four different SOM sizes for a 2x2 dimensional vector, shown in Figure 11.

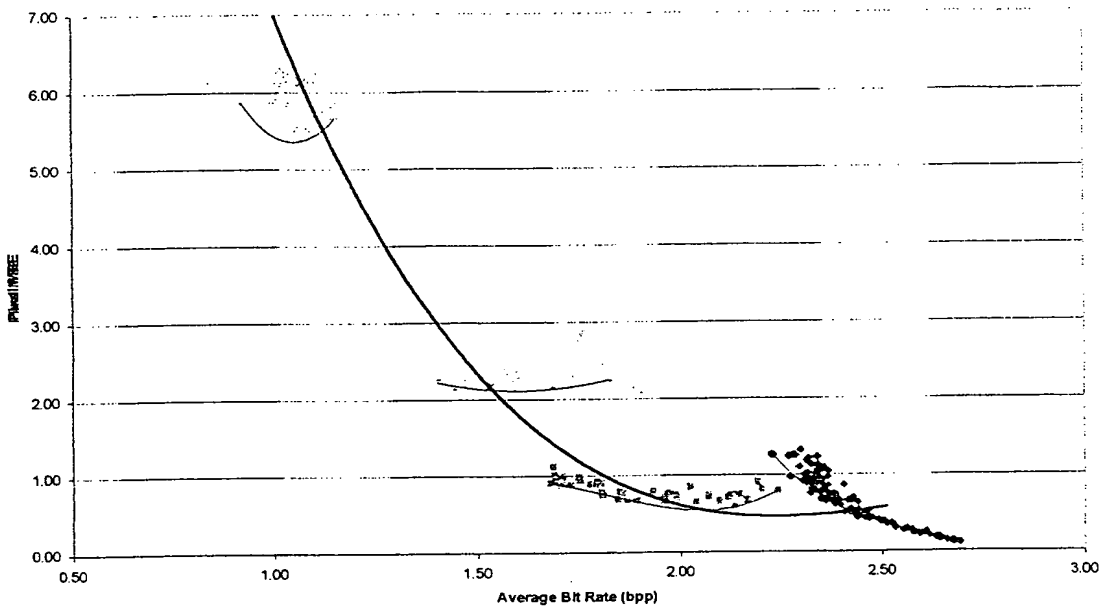


Figure 11: Luminance Level 3 LxHy DWT Sub-Band Rate-Distortion. The ✖'s are the distortion-rate operating points for the SOM size, N=8x8 neurons. The ▲'s are for N=16x16, the ■'s are for N=32x32 and the ◆'s are for N=64x64 neurons.

The overall trendline shows the distortion-rate characteristic for the choice of SOM size. The characteristic could be described as a cost curve where the operating point is chosen depending on slope at that point. A low valued slope (> 1.75 bpp in Figure 11) implies a small distortion cost per coding bit. The large negative slope (< 1.75 bpp) region implies a large bit cost for an improved distortion. From this line, the 32x32 SOM may be chosen as the most optimal under these operating parameters. However, for low bit rate coding the 16x16 SOM will give a gain of ≈0.5 bpp for an average loss of ≈1.0 pixel mean square error (MSE). In this way the operating point is chosen depending on the coding environment.

The thinner trendlines indicate the local convex hull for each SOM. The 64x64 neuron SOM has the typical non-increasing convex shape of high resolution smooth pdf vector quantisers. However, as the number of neurons of the SOM is decreased and therefore the further away

from the high resolution assumptions the vector quantiser operates, two phenomena begin to appear. A clear operational minimum distortion point at a given λ, and multiple operating points for a given average bit rate appear. For the low resolution points it is conceivable that these phenomena may be attributed to a diminishing difference between the global minimum and the local minima for the now large 'volumes of influence' of each neuron in the SOM.

The selected 32x32 SOM shows that the locally optimal operating point is at a minimum. Each sub-band is analysed in the same way to produce an operational SOM size and Lagrange multiplier, λ.

5 DWT-Domain Motion Estimation and Compensation

In spatial domain video codecs (MPEG derivatives, H261 and H263) the primary source of coding gain is from the 'block-based' motion vectors. The underlying principle is to remove the first order temporal redundancy that exists between consecutive frames in a video sequence in an L^2 norm, or in a MSE sense. The image is first divided into n x n pixel blocks. Each block, $I_{nxn,t}$, is overlapped with a block from the reference image (previous predicted image), $R_{nxn,t-1}$, within a pre-defined neighbourhood and the MSE is determined on a pixel by pixel (or half-pixel) basis. In this way the best matching block is chosen as a displacement vector from the same position in the reference image to produce a motion prediction. The motion vector representation is illustrated in Figure 12. The difference between this motion prediction block and the current frame block is calculated as the residual or, as the failure of the prediction. The vector and the residual are coded and sent to the decoder.

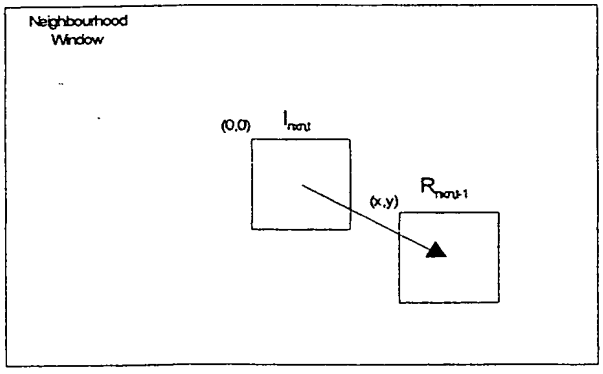


Figure 12: Motion Vector Estimation

Note that the motion estimation and compensation process may be viewed as an adaptive vector quantisation process for vector dimension, k = n x n. The codebook is adapted at each frame and consists of the reference neighbourhood blocks as the codewords. The motion vector is therefore the index into the codebook.

Consider, as a generalisation, the full-pixel full search algorithm for finding the best $\mathbf{L}^2$ norm motion vector in a given neighbourhood. The distortion, D, for the motion vector $\mathbf{v} = [x,y]$ at discrete time, t, may be formulated as:

$$D_{x,y} = \frac{1}{n^2} \sum_{j=0}^{n-1} \sum_{i=0}^{n-1} W(i,j)\left(I_t(i,j) - R_{t-1}(i+x, j+y)\right)^2$$

Here W(i,j) is a weighting function that may be used to 'encourage' the zero vector, or small magnitude vectors, which are usually coded with fewer bits and therefore can act as a rate constraint in the motion vector estimation process. In most applications a rectangular window function is used, W(i,j) = 1. For a neighbourhood window size of ±M, the best matching motion vector is chosen as:

$$\mathbf{v} = \min_{[x,y]}{}^{-1}\left\{D_{x,y}\right\} \quad -M \le x,y \le +M$$

The half-pixel extension to this algorithm requires constructing pixels at half-pixel locations in the reference image. The constructed pixels are bilinear interpolations of the neighbouring pixels. The motion vector search neighbourhood covers the same spatial region but approximates a higher resolution image for a possible lower (long term average) distortion codebook.

The question arises: Can the same coding gain, or at least similar gain, be achieved by applying the same motion estimation and compensation principles to DWT coefficients? From a multiresolution viewpoint, an image is represented by the sum of weighted combinations of approximations to the edges (wavelet bases) at octave resolutions and, a residual texture. Therefore the motion estimation process is a valid prediction process for spatial translational motion, at least in as much as the DWT coefficients represent spatial position. The accuracy of the estimation is dependent on the spatial accuracy of the DWT coefficients bounded by Heisenberg's Uncertainty Principle.

Consider the first iteration of a discrete one-dimensional wavelet transform. The resulting coefficients may be viewed as a decomposition of the signal into a spatio-frequency domain of basis functions. The accuracy of representation of the coefficients is bounded by the

'spread' (uncertainty) of the wavelet function in the spatial domain and the 'spread' of the Fourier transform of the wavelet function in the frequency domain. In the discrete case, the uncertainty may be measured in terms a sampling grid that covers the entire spatio-frequency domain. The grid spacing in the spatial domain, Δx, and the grid spacing in the frequency domain, $\Delta\omega$, is bound by the uncertainty principle (Mallat 1989 and Daubechies 1993):

$$\Delta\omega\Delta x < 2\pi$$

The limit of 2π is the Nyquist rate. Given a discrete image line of X_s pixels ($\Delta x = 1/X_s$), the bound implies that the frequency bandwidth $\Delta\omega < 2\pi/X_s$. Therefore, before any transformation process, the pixel uncertainty is within 1 pixel (normalised to the sampling rate) but the frequency uncertainty of the image line is the entire bandwidth (2π). This is obvious in that no frequency transform has been applied to the image. The implication of this statement is that the accuracy of an optimal motion estimator in the spatial domain is at worst 1 pixel, hence the use of half-pixel estimation. Approximating the wavelet function to an ideal half-band filter and applying the first level of the DWT increases the frequency resolution (or decreases the uncertainty).

$$\frac{1}{2}\left(\frac{2\pi}{X_s}\right)\Delta x < \left(\frac{2\pi}{X_s}\right)$$

The bound limits the spatial uncertainty to $\Delta x < 2$ pixels. Hence an optimal motion estimator operating on the DWT coefficients at the first level, has an inaccuracy of up to 2 pixels. Therefore the normalised spatial domain bound of an ideal DWT at any level, l, may be defined as:

$$\Delta x_l < 2^l \text{ pixels}$$

Convergence to this bound is constrained by practical issues. The DWT is usually implemented with iterative applications of half-band compactly supported perfect reconstruction filter banks with down sampling. These filters have practical limitations in their performance as half-band separators and the down sampling process results in some aliasing. From a frequency domain perspective, small translational motion (<2 pixels) is represented by a phase shift that aliases destructively distorting the DWT coefficients (Nostratinia and Orchard 1995). The filters are not phase invariant and the iterative application of the filters exacerbates the situation such that at higher levels (lower resolutions) translational motion may be added to the coefficients (Villasenor, Belzer and Liao 1995).

Given the uncertainty bound for motion estimation and compensation, there are two possible approaches to operating within the DWT domain.

1. Apply the inverse DWT within the coding loop and motion estimate and compensate in the spatial domain. This approach ensures the maximum motion accuracy but increases the coding complexity (Nosratinia et al. 1995, Houlding and Vaisey 1995).
2. Motion estimate and compensate within the DWT domain and accept the accuracy limitations at each level (Zhang and Zafar 1992). If the MSE of a motion compensated block is greater than that of some weighted value of the pixel energy then, no motion vector is coded (Mandal and Panchanathan 1996). This limits the estimate inaccuracy from increasing the bit rate.

In both approaches advantage may be taken of the multiresolution structure of the DWT for inter level prediction to reduce the general multiresolution redundancy in natural images and reduce the coding complexity. However the partial inverse DWT may be required for the estimation process because there is no direct evidence that the coefficients at lower level sub-bands may be predicted from the higher levels without using the reconstructed LxLy sub-bands.

For sub-band to sub-band DWT domain motion compensation a near optimal solution would be to code a non-zero motion vector only when it out-performs the zero vector in a MSE sense. Following this rule, a 'best-choice' based algorithm is proposed. Within a DWT domain video coding loop, both the pre and post motion compensated vectors are subtracted from the input image coefficients to produce two difference coefficient image vectors. The reconstructed quantised energies of the difference vectors are compared. If the post compensated difference vector energy is less than the pre compensated, then the estimated motion vector is chosen otherwise, the zero vector is selected, with their appropriate quantised difference coefficients, for coding.

6 Video Coding Experiments

Two video test sequences were selected to evaluate the basic and motion compensated algorithms. The sequences are typical for video telephony environments and are of the same subject but at differing distances from the camera. The training difference images for the SOMs were not constructed from any of the test sequences and were taken from a different camera. In this way the generalisation of the algorithm may be tested. The first images of each test sequence are shown in Figure 13 (JillC_10fps.avi) and Figure 14 (JillF_10fps.avi).

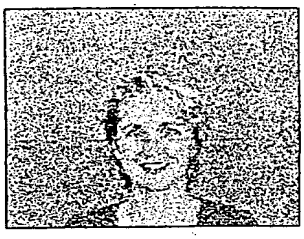


Figure 13: First Frame of Close-in Video Test Sequence

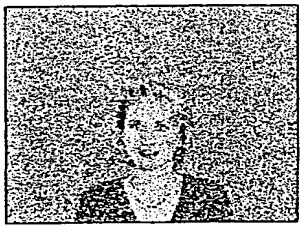


Figure 14: First Frame of Far-out Video Test Sequence

The source test sequences consist of colour images with 24 bits/pixel (bpp) (being 8 bits for red, green and blue components respectively) and with a QCIF pixel resolution (176 x 144). The frame rate is a constant 10 frames/s. For the purposes of comparison, the measure of output quality is considered from a PSNR perspective that is defined from the MSE between the original and the output images on a pixel-by-pixel basis for all colour components. For an input image, I_i, and a reconstructed output image, I_o, with pixel dimensions MxN and C colour components, the MSE is defined as:

$$MSE = \frac{1}{CMN}\sum_{c=1}^{C}\sum_{m=1}^{M}\sum_{n=1}^{N}\left(I_i(c,m,n) - I_o(c,m,n)\right)^2$$

The PSNR is therefore defined as:

$$PSNR = 10\log\frac{255^2}{MSE}\ [dB]$$

For bit rate control a quality factor parameter was used to scale the DWT psychovisual model and a threshold factor was used on the difference image coefficients. An arbitrary quality factor range of q = {0...31} was chosen for the psychovisual scaling, S, applied as a division factor for quantisation and a multiplier for inverse quantisation.

$$S = 1 + \frac{q}{4}$$

The thresholding was applied in the following manner. If the absolute value of the coefficient was less than the threshold value then, it was set to zero. Otherwise, the threshold was

subtracted from (added to) the positive (negative) coefficient. The purpose of the subtraction (addition) was to further 'encourage' small valued coefficient vectors.

The algorithm was applied with target bit rates of 10k bits/s, 28.8k bits/s and 64k bits/s with the quality and threshold factors set according to Table 3.

Table 3: Coding Quality and Threshold Parameters

Test Sequence	Bit Rate (bits/s)	Quality Factor (q)	Threshold
JillC_10fps.avi	10k	3	2
	28.8k	2	2
	64k	1	2
JillF_10fps.avi	10k	3	2
	28.8k	1	2
	64k	1	1

A section of the distortion results from image frame numbers 200 to 300 for the two sequences, comparing the basic algorithm and the extended motion compensated algorithm, for 10k bits/s, 28.8k bits/s and 64k bits/s are shown in Figure 15, Figure 16 and Figure 17 respectfully. The images were decomposed to 4 levels with the DWT but the motion compensation was performed only from level 3.

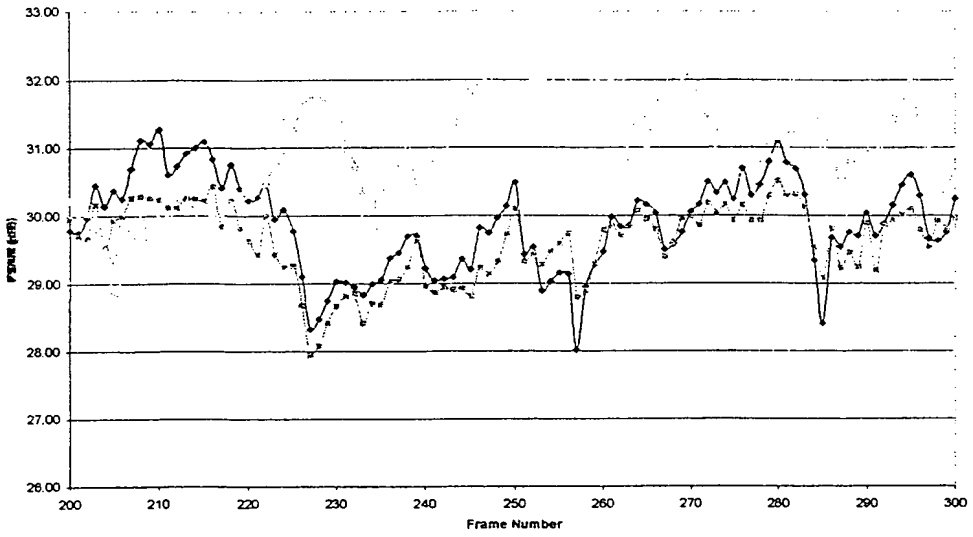


Figure 15: Distortion vs. Video Sequence Frame Number at 10k bits/s. The ✖'s represent the extended algorithm on the JillF_10fps sequence. The ▲'s represent the basic algorithm on the JillF_10fps sequence, the ■'s for the extended algorithm with JillC_10fps and the ◆'s for the basic algorithm with JillC_10fps.

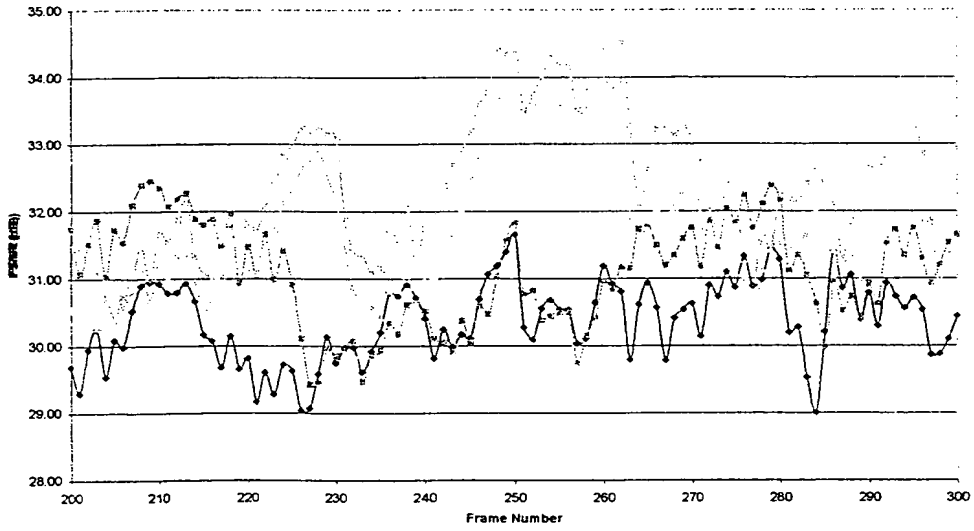


Figure 16: Distortion vs. Video Sequence Frame Number at 28.8k bits/s. The ✖'s represent the extended algorithm on the JillF_10fps sequence. The ▲'s represent the basic algorithm on the JillF_10fps sequence, the ■'s for the extended algorithm with JillC_10fps and the ◆'s for the basic algorithm with JillC_10fps.

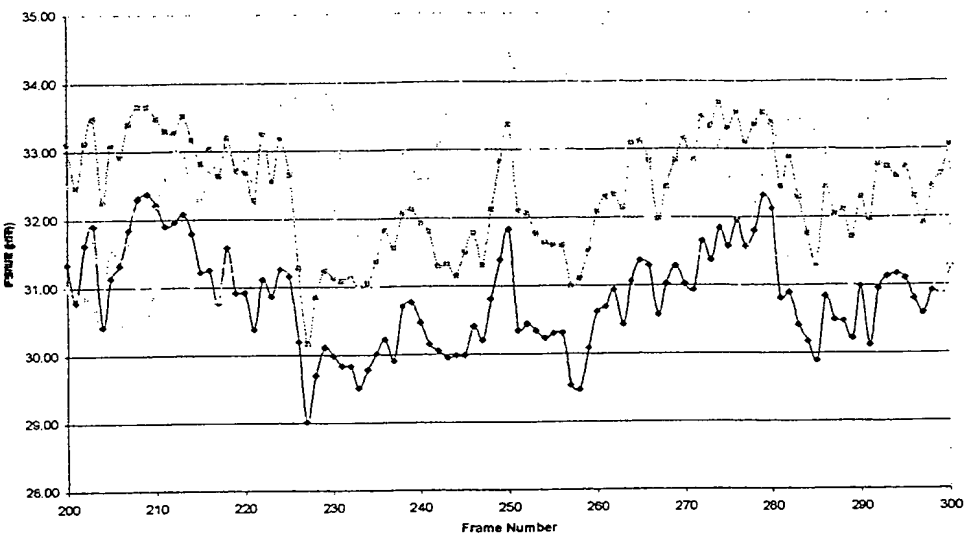


Figure 17: Distortion vs. Video Sequence Frame Number at 64k bits/s. The ✖'s represent the extended algorithm on the JillF_10fps sequence. The ▲'s represent the basic algorithm on the JillF_10fps sequence, the ■'s for the extended algorithm with JillC_10fps and the ◆'s for the basic algorithm with JillC_10fps.

The extended algorithm generally outperforms the basic algorithm but with diminishing returns as the bit rate is decreased from 64k bits/s down to 10kbits/s. The 64k bits/s case provides ≈1.5 dB gain, the 28.8k bits/s case ≈0.5 dB and approximately the same performance at 10k bits/s. The DWT domain motion estimation is more accurate in the high resolution sub-bands than at the low. At very low bit rates the constant frame bit constraint implies that it is likely that the foveated multiresolution coding order will not reach the higher resolutions. Therefore the contribution of the motion compensation to the gain becomes limited and considering that it only begins at DWT level 3. This is more apparent for scenes with higher temporal activity as in the JillC_10fps sequence where the basic algorithm actually performs better. Here, the bit cost of coding the motion vectors outweighs their potential quality gain, although the difference is small.

Note that the sub-band vector quantisers are trained on and hence optimised for difference images that exclude motion compensation. Including motion compensated data in the training process should improve the performance of the extended algorithm.

The effect of the constant frame bit constraint is shown in the shape of the distortion graphs and is consistent with all bit rates. Any sudden temporal activity in the sequence results in a large negative slope of PSNR (≈ frames 230 to 235 of JillF_10fps sequence). There is more

energy in the difference information that requires coding and either more bits must be used and/or the quality must decrease (distortion-rate trade-off). The constant bit constraint implies that the quality is sacrificed. If the activity in the sequence is a burst and low activity follows then, the algorithm will use the 'bit respite' to recover the quality. This is indicated by positive PSNR slope following the decrease (≈ frames 236 to 250 of JillF_10fps sequence). Note that the positive recovery slope is less than the negative quality degradation slope. This degradation and recovery process is the main characteristic of bufferless fixed frame rate algorithms.

7 Conclusion

A statement of the video coding problem has been introduced as a guideline to constructing efficient algorithms. The human interface to the output of coding algorithms implies that an advantage in video information exploitation may be taken if a psychovisual model is included. The acceptability of such algorithms is therefore dependent on a psychological perception analysis. However, this is not within the scope of this research and the output has been evaluated from an objective distortion criterion of PSNR.

A basic and extended algorithm is proposed as a bit rate scalable solution to video telephony applications. A fixed frame rate of 10 frames/s and bufferless operation were constraints placed on the algorithm for very low bit 'live' operation. The psychovisual model applied to the algorithms included operating in the DWT domain, spatial frequency shaping of the coefficients and a multiresolution foveated sub-band coding order. The frame difference coefficients were vector quantised and variable length encoded. The extended algorithm added DWT domain motion estimation and compensation to the basic algorithm. A 'best choice' algorithm was developed to optimise the performance of the motion compensation process. The performance of the algorithms were compared with one another at 10k bits/s, 28.8k bits/s and 64k bits/s. The result indicates that the added complexity of the extended algorithm provides a distortion gain ≈ 1.5 dB at 64k bits/s but diminishes towards the very low bit rate range at 10k bits/s.

A spatial DWT quantisation table was established based on the visibility threshold of quantisation errors within each sub-band as a function of spatial frequency, level, orientation and colour. The table was scaled with a linearising function to provide a method of algorithm bit control. The vector quantisers for each sub-band were constructed choosing an operational point on a distortion-rate function measured for the corresponding sub-band. They were trained with a modified rate-constrained SOM algorithm on DWT difference coefficients. The

modification adapted the training data to compensate for operation within the first order coding loop. Heisenberg's Uncertainty Principle limits the effectiveness of DWT domain block based motion compensation. The limitations are taken into account by the 'best choice' algorithm and confining the levels at which the compensation is applied.

References

Antonini M., Barlaud M., Mathieu P., Debechies I. *Image Coding Using Wavelet Transform.* IEEE Transactions on Image Processing, Vol. 1, No. 2, April 1992, pp. 205 – 220.

Campbell F., Kulikowski J. *Orientation selectivity of the human visual system.* Journal of Physiology, Vol. 197, 1966, pp. 437 – 441.

Campbell F., Robson J. *Application of Fourier analysis to the visibility of gratings.* Journal of Physiology, Vol. 197, 1968, pp. 551 – 566.

Chou P.A., Lookabaugh T., Gray R.M. *Entropy-Constrained Vector Quantization.* IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 37, January 1983, pp. 31 – 42.

Daubechies I. (Editor). Different Perspectives on Wavelets. American Mathematical Society, 1993. *Wavelet Transforms and Orthonormal Bases.* Proceedings of Symposia in Applied Mathematics, Vol. 47, American Mathematical Society Short Course, Texas, January 1993, pp. 1 – 33.

Gersho A., Gray R.M. Vector Quantization and Signal Compression. Kluwer Academic Publishers, Boston, 1992.

Haykin S. Neural Networks A Comprehensive Foundation. Macmillan College Publishing Company, New York, 1994.

Houlding D., Vaisey J. *Pyramid Decompositions and Hierarchical Motion Compensation.* SPIE Proceedings on Digital Video Compression: Algorithms and Technologies, Vol. 2419, 1995, pp. 201 – 209.

International Telecommunication Union. *VIDEO CODING FOR LOW BIT RATE COMMUNICATION.* TRANSMISSION OF NON-TELEPHONE SIGNALS. ITU-T TELECOMMUNICATION STANDARDIZATION SECTOR OF ITU. ITU-T Recommendation H.263, 1996.

Kohonen T. Self-Organizing Maps. Springer-Verlag, New York, 2nd Edition, 1997.

Linde Y., Buzo A., Gray R.M. *AnAlgorithm for Vector Quantizer Design.* IEEE Transactions on Communications, Vol. 28, January 1980, pp. 84 – 95.

Luttrell S.P. *Self-organization: A derivation from first principle of a class of learning algorithms.* IEEE Conference on Neural Networks, Washington, DC, 1989, pp. 495 – 498.

Mallat S.G. *Multifrequency Channel Decompositions of Images and Wavelet Models.* IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 37, No. 12, December 1989, pp. 2091 – 2110.

Mandal M.K., Panchanathan S. *Motion Estimation Techniques for a Wavelet-based Video Coder.* SPIE Proceedings on Digital Video Compression: Algorithms and Technologies, Vol. 2668, 1996, pp. 122 – 128.

McLaren D.L., Nguyen D.T. *Removal of subjective redundancy from DCT-coded images.* IEE Proceedings I, Vol. 138, No. 5, October 1991, pp. 345 – 350.

Ngan K.N., Leong K.S., Singh H. *Adaptive cosine transform coding of images in perceptual domain.* IEEE transactions on Acoustics, Speech, and Signal Processing, Vol. 37, No. 11, 1989, pp. 553 – 559.

Nosratinia A., Orchard M.T. *A Multi-Resolution Framework for Backward Motion Compensation.* SPIE Proceedings on Digital Video Compression: Algorithms and Technologies, Vol. 2419, 1995, pp. 190 – 200.

Sakrison D.J. *On the role of the observer and a distortion measure in image transmission.* IEEE transactions on Communications, Vol. 25, No. 11, 1977, pp. 1251 – 1267.

Sweldens W. *The Lifting Scheme: A Custom-Design Construction of Biorthogonal Wavelets.* Transactions on Applied and Computational Harmonic Analysis, Vol. 3, No. 2, April 1996, pp. 186 – 200.

Van den Branden Lambrecht C.J., Verscheure O. *Perceptual Quality Measure using a Spatio-Temporal Model of the Human Visual System.* SPIE Proceedings on Digital Video Compression: Algorithms and Technologies, Vol. 2668, 1996, pp. 450 – 461.

Villasenor J.D., Belzer B., Liao J. *Wavelet Filter Evaluation for Image Compression.* IEEE Transactions on Image Processing, Vol. 4, No. 8, August 1995, pp. 1053 – 1060.

Watson A.B., Yang G.Y., Solomon J.A., Villasenor J. *Visual Thresholds for Wavelet Quantization Error.* SPIE Proceedings on Human Vision and Electronic Imaging, B. Rogowitz and J. Allebach, Editors., Vol. 2657, Paper No. 44, 1996.

Zhang Y.Q., Zafar S. *Motion-Compensated Wavelet Transform Coding for Color Video Compression.* IEEE Transactions on Circuits and Systems for Video Technology, Vol. 2, No. 3, September 1992, pp. 285 – 296.